DOCUMENT RESUME

ED 030 148 CG 003 379

By-Stanley, Julian C.; Wang, Marilyn D.

Differential Weighting: A Survey of Methods and Empirical Studies.

Johns Hopkins Univ., Baltimore, Md. Center for the Study of Social Organization of Schools.

Spons Agency-College Entrance Examination Board, New York, N.Y.

Pub Date Nov 68

Note-82p.

EDRS Price MF-$0.50 HC-$4.20

Descriptors-Correlation, Literature Reviews, *Measurement Instruments, *Measurement Techniques, Research, *Scientific Concepts, *Statistical Analysis, *Testing, Test Interpretation

The literature on a priori and empirical weighting of test items and test-item options is reviewed. While multiple regression is the best known technique for deriving fixed empirical weights for component variables (such as tests and test items), other methods allow one to derive weights which equalize the effective weights of the component variables (their individual contributions to the variance of the composite). Fixed weighting is most effective, in general, when there are few variables in the composite, and when these variables are not highly correlated. Variable weighting methods are those in which there is no nominal weight, constant over subjects, applied to a single item or response option. Of most interest are variable response-weighting methods such as those recently suggested by de Finetti (1965) and others. To be effective, such weighting methods require that the subject be able to maximize his expected score only if he reports his subjective probabilities honestly. Variable response-weighting methods, perhaps in conjunction with fixed response-weighting methods, show promise for increasing the reliability and validity of test scores. (Author/CJ)

differential weighting

A Survey of Methods and Empirical Studies

JULIAN C. STANLEY
Professor of Education
and Psychology and
Head, Experimental Design
and Statistics Analysis Unit,
Center for the Study of
Social Organization of
Schools,
The Johns Hopkins University,
Baltimore/ Maryland/21218
Principal Investigator

MARILYN D. WANG
Research Associate
Department of Psychology
The Johns Hopkins University
Co-Investigator

Supported by the
College Entrance
Examination Board,
475 Riverside Drive,
New York/ New York/ 10027

November/ 1968

Contents

Abstract

When a number of psychological measures are to be combined it is sometimes desirable to weight the measures differentially, either with fixed weights which are constant for all subjects or with variable weights which are not. In this paper we review the literature on *a priori* and empirical weighting of test items and test-item options.

A large number of methods are available for deriving fixed empirical weights for component variables such as tests and test items. The best known and most widely used technique is multiple regression. Other methods allow one to derive weights which equalize the effective weights of the component variables, *i.e.*, their individual contributions to the variance of the composite, or which equalize the correlation of each variable with the composite, or which maximize composite reliability. Other weighting methods which have been popular include weighting by the reciprocal of the standard deviation, weighting (tests) by length or difficulty, and weighting by the validity coefficient of the component variable.

The effectiveness of fixed weighting depends on the number of measures to be combined, their intercorrelations, and certain characteristics of the weights. In general, fixed weighting is most effective when there are few variables in the composite and when these variables are not highly correlated. For a largenumber of positively correlated variables (such as test items) the correlation between two randomly weighted composites rapidly approaches unity.

Fixed weighting has also been used to develop scores for response categories such as those in the items of personality, attitude, and interest tests, where there is no "correct" response option. The raw data in such cases are classificatory rather than quantitative. A familiar example of one such

method is that used by E.K. Strong Jr. to secure option weights for the Strong Vocational Interest Blank.

Empirical studies of fixed weighting, popular in the 1920's and 1930's, demonstrated what the analytical papers predicted would be the case. Weighting was found useful in many cases where only a few tests were combined in a battery. But weighting the items of a long test was shown repeatedly to be ineffective, or so slightly effective as to be impractical. There are few empirical studies of response-option weighting in achievement or aptitude tests, although there is reason to believe that such weighting might be effective despite the fact that item weighting is not.

Variable weighting methods are those in which there is no nominal weight, constant over subjects, applied to a single item or response option. Of most interest are variable response-weighting methods such as those recently suggested by de Finetti (1965) and others. Here, the subject's response to a test item need not be restricted to simply selecting a single response option as correct. Rather, he may be asked to respond in one of a variety of ways. In particular, he may be instructed to assign a probability to each response option corresponding to his subjective probability of the correctness of the option. A scoring formula is then used to take the probability distribution into account in arriving at a score for the item. To be effective, such weighting methods require that the subject be able to maximize his expected score if and only if he reports his subjective probabilities honestly.

The de Finetti subjective-probability approach does not produce differential scoring weights for the various distracters, however, nor do the methods devised by Birnbaum and by Cleary. A criterion-keying procedure due to Guttman does provide differential scoring weights for the various options of a multiple-choice item and seems promising enough to be tried, now that high-speed digital computers are readily available.

Variable response-weighting methods, perhaps in conjunction with fixed response-weighting methods, show promise for increasing the reliability and validity of test scores, a feat which cannot be attained with fixed item-weighting techniques for long tests composed of positively intercorrelated items.*

* * * * * *

*For a shorter version of this review that is considerably more detailed than this abstract, see Stanley & Wang (1968).

Whenever several measures are to be combined to form a single composite measure or to predict a criterion, the question of differential weighting of the component measures presents itself. Can differential weighting improve the reliability of measurement and/or provide a more valid composite measure than would be obtained if the component measures were merely summed or averaged?

Theoretically, the answer to this question should be "Yes" for both reliability and validity. It is unlikely that all of the component measures will be equally reliable, have equal variances, be equally intercorrelated with one another, and be equally correlated with the underlying variable which the composite is supposed to measure or with the external criterion. But all of these characteristics of the component measures will be reflected in the composite measure. Thus, on purely logical grounds, it is to be expected that differential weighting would be effective.

If criterion measures are available, multiple-regression techniques will provide a set of weights which is optimal for minimizing error of prediction for the group on which the weights were derived, under the usual assumptions of normality and linearity of regression. When no external criterion is available, certain assumptions concerning the nature of the variable which the composite is supposed to measure enable us to identify those component measures which should be weighted more heavily. Or, alternatively, weights may be chosen so as to maximize certain internal criteria such as the reliability of the composite measure. Regardless of which method is used to derive the weights, however, all methods have in common the fact that they weight most heavily those measures which are "best" according to the criterion adopted in each particular instance, and they weight least, perhaps even negatively, those measures which are worst.

McDonald (1968) has offered "a unified treatment of the weighting problem," a general procedure for obtaining weighted linear combinations of variables. This general procedure includes as special cases multiple-regression weights, canonical variate analysis, principal components, maximizing composite reliability, canonical factor analysis, and some other well known methods. He shows that the general procedure yields certain desirable invariance properties with respect to transformations of the variables. McDonald's approach is applicable to a considerable part of this survey, because it undoubtedly can be used to simplify some of the seemingly diverse procedures of the past half century.

Although differential weighting promises, in theory, to provide substantial gains in predictive or construct validity, very often, in practice, these gains are so slight that they do not seem to justify the labor involved in deriving the weights and scoring with them. This is especially true when the component measures are a large number of test items and much less true when they are a small number of tests comprising a battery. It is this fact which has led psychologists to conclude that, in general, weighting is not worth the trouble, especially as far as *item* weighting is concerned. (For example, see Guilford, 1954; Gulliksen, 1950.)

But item weighting is not the only type of weighting which has been investigated. Multiple regression is very often effective when a team of variables, not necessarily tests, is used to predict a criterion. In most interest and personality tests some form of *option* weighting occurs, *i.e.*, the subject's score on a given item depends on which option he selects or prefers. In this case it is the options which are differentially weighted. Usually there are many sets of weights which are applied successively to the answer sheet in order to derive a score for the subject on a number of different scales. Although it has not been studied extensively in the past, differential weighting of item options on academic aptitude or achievement tests has also been considered a possibility. In fact, it has recently been proposed (de Finetti, 1965; Shuford, Albert & Massengill, 1966) that the reliability and validity of tests may be increased if the subject himself assigns weights to the options according to his confidence in the correctness of each option.

The remainder of this paper will be devoted to a systematic study of the weighting question. First, different types of weighting and methods of deriving weights will be discussed, as well as the mathematical restrictions which limit the effectiveness of certain sets of weights, regardless of what method is used to derive them. Next, a summary of empirical investigations of weighting in each of the typical situations where weighting has been considered potentially useful will be presented. Finally, consideration will be given to the recently suggested confidence-weighing methods.

Weighting and the Derivation of Weights

In this and the following two sections we shall be concerned only with what will be termed *fixed* weighting. In this context "fixed" implies that the weight for a given measure is constant for all individuals. Thus, if the items of a single test are differentially weighted, the same set of weights is used for all examinees. In contrast, we will term *variable* weighting all methods where the weights are free to vary from person to person. Variable weighting will be discussed in a later section.

A Definition of Weighting

It is customary to define the weight of a single variable in a composite in terms of the contribution of that variable to the variance of the composite. The contribution of each of n component variables to the variance of the composite is equal to the variance of the variable plus the $n-1$ covariances of that variable with the $n-1$ other variables in the composite. This follows directly from the formula for the variance of a sum:

$$(1) \qquad \mathrm{Var}(X_1 + X_2 + \ldots + X_n) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \ldots + \mathrm{Var}(X_n) + 2\mathrm{Cov}(X_1X_2) + 2\mathrm{Cov}(X_1X_3) + \ldots + 2\mathrm{Cov}(X_{n-1}X_n).$$

The formula indicates that the variance of the composite is equal to the sum of n variance terms and $n(n-1)$ covariance terms corresponding to the $n(n-1)$ combinations and permutations of pairs of the n variables. If the variances and covariances are arranged in a symmetrical matrix of order n x n, the contribution of the *i*th variable to the variance of the composite is given by the sum of the terms in the *i*th row or the *i*th column of the matrix. Thus, the n variables which comprise the composite are equally weighted if and only if they make equal contributions to the total variance, *i.e.*, the sum of the elements in each row (or column) of the variance-covariance matrix is equal to a constant.

Although it is seldom stated, this definition of weighting implies that the resulting composite measure, for a single individual, has little significance in and of itself and that its meaning is derived via the total distribution of the composite measures for all individuals. This is probably not an unreasonable assumption, since so much measurement in psychological research is of the ordinal or interval variety and population norms of some kind are required for interpretation of a single individual's score. However, in certain cases, the composite and component measures do have sufficient intrinsic or arbitrarily agreed-upon meaning that the score for an individual may be interpreted without reference to a distribution of scores. When this is the case, it will be shown that the above definition of weighting is not appropriate.

Nominal vs. Effective Weighting. The approach outlined above may also be followed to determine the contribution to the total variance of each of a number of variables which have been "weighted" before being summed. Assume a set of n variables X_i $(i = 1,\ldots,n)$, and a corresponding set of weights w_i $(i = 1,\ldots,n)$, such that the composite score for any individual is given by $w_1X_1 + w_2X_2 +\ldots+ w_nX_n$. The variance of the composite is given by

$$(2) \qquad \mathrm{Var}(w_1X_1 + w_2X_2 +\ldots+ w_nX_n) =$$

$$w_1^2\mathrm{Var}(X_1) + w_2^2\mathrm{Var}(X_2) +\ldots+ w_n^2\mathrm{Var}(X_n) + 2w_1w_2\mathrm{Cov}(X_1X_2)$$

$$+ 2w_1w_3\mathrm{Cov}(X_1X_3) +\ldots+ 2w_{n-1}w_n\mathrm{Cov}(X_{n-1}X_n).$$

Again, the contribution of any one variable to the variance of the composite is given by the sum of the elements in the corresponding row (or column) of the variance-covariance matrix. The w_i's constitute the *nominal weights*, whereas the *effective weight* of each variable is defined in terms of its contribution to the total variance of the composite.

When nominal weights are unity the contribution of the ith variable to the total variance of the composite, C_i, is given by

$$(3) \qquad C_i = \mathrm{Cov}(X_1X_i) + \mathrm{Cov}(X_2X_i) +\ldots+ \mathrm{Cov}(X_{i-1}X_i) + \mathrm{Var}(X_i) + \mathrm{Cov}(X_{i+1}X_i)$$

$$+\ldots+ \mathrm{Cov}(X_nX_i).$$

The natural effective weight of the ith variable is thus given by its own variance plus its covariance with each of the remaining variables, or equivalently, by its covariance with the sum of the remaining variables.

When nominal weights w_i are assigned to the variables, the expression for the contribution of the ith variable to the variance of the composite is given by

$$(4) \qquad C_i = w_1 w_i \mathrm{Cov}(X_1 X_i) + w_2 w_i \mathrm{Cov}(X_2 X_i) + \ldots + w_{i-1} w_i \mathrm{Cov}(X_{i-1} X_i)$$

$$+ w_i^2 \mathrm{Var}(X_i) + w_{i+1} w_i \mathrm{Cov}(X_{i+1} X_i) + \ldots + w_n w_i \mathrm{Cov}(X_n X_i).$$

It is a common misconception that when nominal weights have been assigned they correspond to the relative weights of the variables in the composite. Equation (4) indicates that this is not the case. Although the nominal weights do influence the effective weights, they are not in general proportional to them.

Equation (4) may also be expressed in terms of the intercorrelations of the variables. Since $\mathrm{Cov}(X_1 X_i) = r_{1i} s_1 s_i$, the formula becomes

$$(5) \qquad C_i = w_1 w_i r_{1i} s_1 s_i + w_2 w_i r_{2i} s_2 s_i + \ldots + w_{i-1} w_i r_{i-1,i} s_{i-1} s_i + w_i^2 s_i^2$$

$$+ w_{i+1} w_i r_{i+1,i} s_{i+1} s_i + \ldots + w_n w_i r_{ni} s_n s_i.$$

From this it may be seen that the contribution of the *i*th variable to the total variance depends on (a) the nominal weights w_i, (b) the variance of X_i, (c) the $n-1$ correlations between X_i and the $n-1$ other variables in the composite, and (d) the standard deviations of the other $n-1$ variables.

Now assume that each of the n variables is given in standard form, $\zeta_i = (X_i - \mu_i)/\sigma_i$, or equivalently, that each X_i is divided by the appropriate standard deviation σ_i. In this case all variances and standard deviations of the resulting scores will be equal to unity and therefore will disappear from the formula, giving

$$(6) \qquad C_i = w_1 w_i r_{1i} + w_2 w_i r_{2i} + \ldots + w_{i-1} w_i r_{i-1,i} + w_i^2 + w_{i+1} w_i r_{i+1,i}$$

$$+ \ldots + w_n w_i r_{ni}.$$

Thus, when scores are expressed in standard form the effective weight of the *i*th variable is determined by the nominal weights and the intercorrelations of the variables. If unit weights are used with standard variables, the effective weight of a variable is approximately proportional to its average correlation with the other variables:

$$C_i = 1 + \sum_{j=1}^{n} r_{ij} \qquad (i \neq j) \tag{7}$$

$$= 1 + (n - 1)\bar{r}_{ij}$$

$\sum_{j=1}^{n} r_{ij}$ is the correlation of the ith variable with the total score on the remaining variables, $r_{X_i, \sum_{j=1}^{n} X_j}$.

From the foregoing discussion two things should be clear. First, the nominal weights will not in general be proportional to the effective weights. Second, only rarely will variables have equal effective weights unless the nominal weights have been derived specifically to ensure this result (*e.g.*, see Kaiser, 1967, for a way to make all Cov(X_iX_j) zero). Otherwise, using unit nominal weights with standard scores probably comes closest to achieving this end, particularly if the average correlation of each variable with the others is nearly constant.

An Exception. There is one situation, however, in which the nominal weights are *always* directly proportional to the effective weights. This is the situation alluded to earlier where the usual definition of effective weighting is not appropriate.

Assume, for example, that a teacher decides in advance to assign grades to her class on the basis of a "semester score" which is expressed as a percentage. The following scheme might be adopted: A = 90%-100%; B = 80%-89%; C = 70%-79%; D = 65%-69%; F = below 65%. Five examinations[1] are given

[1] If the examinations are of equal length, the following applies to the items as well as to the test as a whole.

during the semester and the score on each is expressed as the percentage of items answered correctly. The semester score is the arithmetic average of the five examination scores, and the final letter grade is assigned according to the predetermined scheme. In this case it is appropriate to say that the five examinations have been equally weighted in determining the final grade regardless of the distribution of scores on any of the examinations or the intercorrelations of the examinations. Likewise, if a *weighted* average of the examinations had been taken, the effective weight of each would be directly proportional to the nominal weight assigned to it. This is true because, *in this case,* the semester score of each pupil may be interpreted by itself, with no reference to the distribution of semester scores of which it is a part. Since the pupil's semester score will be interpreted directly, *i.e.,* assigned a letter grade, and since the several examinations contribute to this score in direct proportion to the weights assigned to them, these nominal weights are also the effective weights.

This point can be seen even more clearly at the item level. Suppose that the teacher administers a five-question test and assigns 35 percentage points to one of the questions. No pupil who receives 0 percentage points on that question earns a grade of C or better. If all the pupils fail the question completely, they all earn grades of D or F, even though scores on the question have 0 variance and covary 0 with scores on each other question. Absolute grading on an arbitrary scale differs in this way from grading each pupil relative to the performance of the other pupils in his class.

This situation is to be contrasted with that in which the examinee's semester score is interpreted with reference to the total distribution of such scores. Suppose, instead, that the letter grades were to be assigned on the basis of the examinee's standard semester score, according to the scheme in Figure 1.

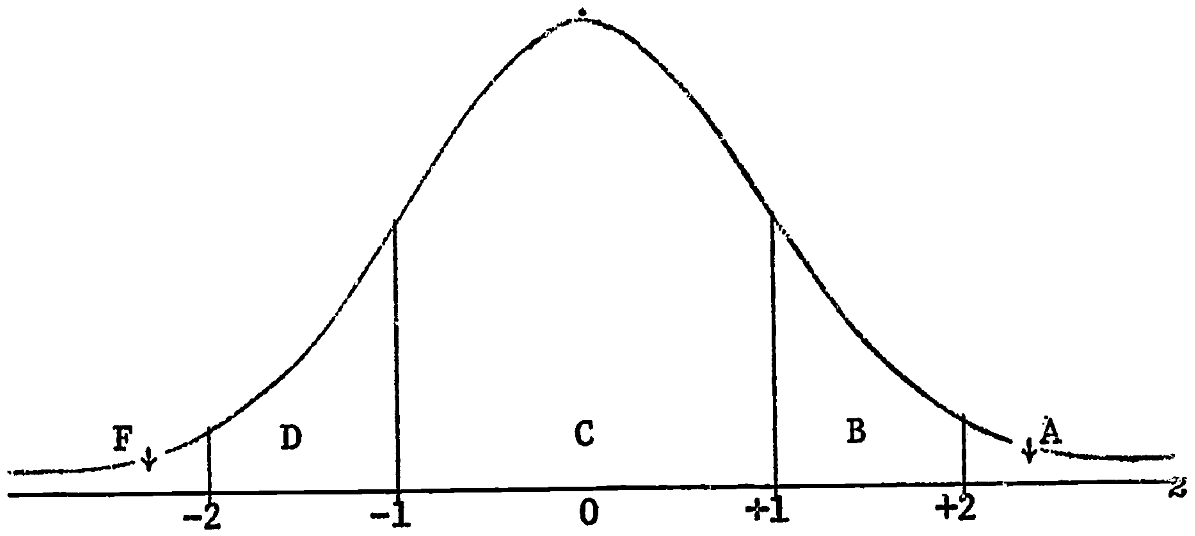

Figure 1. Grading system using standard scores.

In this case the examinee's letter grade depends both on his own score and on the variance of the semester-score distribution. It is for this reason that the effective weight of the several examinations is assessed via the contribution-to-variance criterion.

Although the latter method of assigning grades, or some modification of it, is very common, particularly at the college level, the former method is probably sufficiently common to account for the intuitive feeling of many that the nominal weights are indeed the effective weights. As suggested earlier, however, aside from the classroom situation, the former type of measurement is sufficiently rare in psychology to justify the adoption of the contribution-to-variance definition of effective weighting.

Methods of Weighting Variables

We are now ready to consider in greater detail the specific methods of weighting which have been and continue to be used in psychological research. In each method the entity to be weighted is a quantitative variable, in contrast to methods discussed in a later section where it is unordered response categories for which scoring weights are sought. For the most part, the methods of this section are concerned with assigning weights to tests in a battery or to test items. In most of the methods the weights derived are the nominal weights, *i.e.*, the multiplicative constants by which the measures on the *n* component variables are weighted. In some cases derivations and formulas assume considerably simpler forms if the measures are expressed in a standard form rather than in raw-score from. In all such cases it is implicit that, if desired, the derived weights may be redefined to absorb the standard deviations of the variables.

Random Weights. When raw scores on a number of variables are simply summed or averaged to form a composite measure the effective weight of each variable is determined by Equation (3). Since no deliberate effort is made to control the effective weights of the variables they will be termed *random weights*. The term "random" should not be taken to indicate that differences between the effective weights are due to "chance." Real differences in the variances of the component variables and differences in their intercorrelations are simply allowed free rein in determining the weights. It should be carefully noted that this case corresponds to what is usually call "no" weighting. It must be remembered that these measures are unweighted only in the sense that the nominal weights are unity.

A Priori Weights. When nominal weights are assigned to the *n* component variables on the basis of judgments or ratings or some similar procedure, the weights are termed *a priori weights*. The decision not to weight, *i.e.*, to assign unit nominal weights, is a special case of this.

A very common case of *a priori* weighting occurs when different sections or questions on an examination are weighted differentially. For example, 20 true-false items on a test might be allowed a point apiece, whereas 20 multiple-choice questions on the same test may be worth 4 points apiece. In some cases even items of the same type may be differentially weighted on an *a priori* basis. Corey (1930) had instructors rate each item of an objective test in psychology on a 7-point scale according to its judged importance for a general knowledge of psychology. The rating then became the weight for the item. Weighting on an *a priori* basis is also very common in personnel decisions, where certain job criteria may be deemed more important than others.

Although there are important empirical methods available for deriving nominal weights, this should not be taken to mean that such methods are necessarily preferable in all situations. Burt (1950, p.122) concludes that *a priori* or subjective weighting may be necessary where questions of value are concerned or where the criterion is genuinely composite.

Empirically Derived Weights. Of all of the empirical methods of deriving predictor weights, the one probably most familiar to psychologists is multiple (linear) regression. This is but one of a number of least-squares solutions which have been used to derive weights. The other methods have proved extremely useful since it is so often difficult to find an adequate criterion variable. These methods will be discussed in turn and their major advantages noted. Since the actual mathematical derivations of the weights are available elsewhere they will not be presented here.

Multiple Regression. If for a certain population measures on each of n predictor variables X_i are available together with measures on the variable to be predicted, X_0, the classical multiple regression equation will give the optimal weights to be assigned to the predictors in order to maximize the correlation between the predicted or composite score and the actual criterion score. This solution also minimizes the mean squared error of prediction, given that the function expressing the relationship between the predictors and the criterion is linear.

The general form of the equation when all variables are expressed in raw-score form is

$$(8) \qquad \hat{X}_0 - \mu_0 = b_{01.23\ldots n}(X_1 - \mu_1) + b_{02.13\ldots n}(X_2 - \mu_2) + \ldots + b_{0n.12\ldots n-1}(X_n - \mu_n),$$

where $\hat{X}_0$ is the predicted criterial score for an examinee, μ_0 is the population mean on the criterial variable, and the b's are population weights for deviation-from-the-mean predictor scores, $(X_i - \mu_i)$.

This equation can be simplified by expressing all predictors in standard form and the predicted score in semi-standard form, as follows:

$$(9) \qquad (\hat{X}_0 - \mu_0)/\sigma_0 = \beta_{01.23\ldots n}z_1 + \beta_{02.13\ldots n}z_2 + \ldots + \beta_{0n.12\ldots n-1}z_n.$$

The b-weights in Equation (8) are the nominal weights for scores used to predict the criterion score. The β-weights in Equation (9) are the nominal weights for standard scores. They are related to the b-weights by the equation

$$(10) \qquad b_{0i.12\ldots n} = \beta_{0i.12\ldots n}(\sigma_0/\sigma_i).$$

$\beta_{0i.12\ldots n}$ is the partial regression coefficient of X_0 on X_i. Specifically, it is the regression of that part of X_0 which is independent of all the other n-1 variables on that part of X_i which is also independent of them (Kelley, 1923).

Note the two following properties of the regression weights *ceteris paribus:*

1. The larger the correlation between the variable and the criterion, the larger the weight.
2. The more independent the variable of the n-1 other variables, the larger the weight.

In the equations above, nothing has been assumed concerning the source of the criterion variable except that measures are available in the population of interest. Ryans (1954) has considered the problem of weighting from the other side of the fence, *i.e.*, weighting the components of the criterion to arrive at a suitable measure. Hotelling (1935) has presented a method called "canonical correlation" for assigning weights to *two* batteries (one of which might serve to define a criterion) so as to maximize the correlation between them.

A word of caution is in order concerning the use of multiple regression. The weights derived via multiple regression are the weights which maximize the multiple correlation between predictors and criterion for the particular set of measures on which the weights are derived. This is true whether the set of measures is from an entire population or merely a sample from a population. Most often, however, the weights are derived on a sample and then subsequently used to predict the criterion in the entire population. The multiple correlation between actual criterion scores in the population and criterion scores predicted via sample weights will necessarily be less than the multiple correlation which *could* be obtained if scores for the entire population were used to derive the weights. However, if the weights are derived on a random sample from the population, then the observed values of intercorrelations and component variances are likely to be representative of the values of the population parameters,and the obtained weights are likely to be reasonable approximations to the optimal weights for the population. If this is the case, then the multiple correlation obtained using sample weights in the population should not be much less than the maximum possible multiple correlation for the population. Quite commonly, however, the multiple correlation obtained from using sample weights in the population is not only less than the maximum correlation but also less than the sample multiple correlation. This "shrinkage" in the multiple correlation has received considerable attention in the psychological literature.

When there is error of measurement in either the predictors or in the criterion, shrinkage of the multiple correlation may be even more dramatic. In the previous example the sample multiple correlation could be said to roughly indicate the predictability of the criterion. More exactly, the square of the multiple correlation corresponds to the proportion of the criterion variance which can be "explained" by the predictors. It was assumed that the measures themselves were error-free. If, however, the predictors are not error-free measures, then a certain proportion of the "predicted" or "explained" variance is in actuality error variance which is random from sample to sample. However, in any given sample from the population error will affect the value f the *r*'s and *s*'s on which the multiple regression weights are based. Within the sample, the weights are actually tailored to "predict" both error variance and "true" variance. Thus, in this case there are two factors which affect the representativeness of the *r*'s and *s*'s, sampling fluctuations and measurement error. Thus, when there is measurement error, weights derived from a sample are more likely to deviate from optimal weights than when there is no measurement error. In this case the optimal weights would be those which would be obtained if error-free population measures were available. Unless the measuring instrument can be made error-free, these weights can never be known. The important point here is that if the ultimate goal of the regression analysis is prediction, then the presence of unwanted random-error variance in the sample increases the likelihood that *r*'s and *s*'s obtained in the sample will not be representative of the values of the corresponding parameters in the population. Thus, when sample regression weights are based on fallible measures, it is extremely important to crossvalidate the weights before reporting validity coefficients.

The error of measurement problem with respect to regression weights is most apparent when the predictors are psychological tests, although tests are by no means the only psychological measures subject to error! Wolins (1967) has discussed problems of hypothesis testing and estimation for this case, noting that as the intercorrelations between the predictors rises, bias in the regression coefficients also rises.[2] This is due to the fact that as the correlation between two fallible variables approaches the limit set by their respective reliabilities, the differences between the variables increasingly reflects error variance rather than true variance. As the intercorrelation between the variables drops there is less bias in the weights since proportionally more of the difference between the variables is due to "true" differences rather than error. The estimate of the multiple correlation squared, however, is only slightly affected by bias in the regression coefficients, since as the intercorrelation rises it is less and less dependent on the actual values of the regression coefficients.

In terms of efficiency, multiple-regression techniques will be most useful when there are but few predictor variables, and, as the number of predictor variables rises, when the predictors are relatively independent. (Suppressor variables constitute an exception to this rule, since they increase the multiple correlation as their correlation with the other predictors *rises*.)

Equal Contributions to Total Variance. It is sometimes desirable to ensure that each of the component variables has equal effective weighting. This might be the case if the composite measure is truly intended to be a composite rather than a measure of some hypothesized underlying unitary entity, as when each of *n* judges assigns ratings and it is desired that the judges' opinions have equal weight. Or, in the absence of an external criterion, equal effective weighting may be deemed appropriate.

[2] Cureton (1951) has also discussed this problem.

Both Wilks (1935) and Dunnette & Hogatt (1957) have presented iterative procedures for obtaining approximate solutions for this case. The solution requires setting n equations like (5) equal to an arbitrary constant and solving for the weights w_i. If, rather than equal weighting, it is desired to set the weights in some predetermined proportions to one another, this may be accomplished by setting the constants in that proportion. Thus, the method allows us to assign *a priori* effective weights rather than *a priori* nominal weights as discussed in a previous section.

An interesting special case of this method occurs when uncorrelated standard scores are given unit (or *a priori*) weights. (See Kaiser, 1967, for a rather interesting orthogonalizing procedure.) Only under this condition does the use of standard scores ensure equal weighting.

Equal Correlations with the Composite. When there is no external criterion available, weights may be derived by the method of least squares to equalize the correlation of each variable with the resulting weighted composite score (Wilks, 1938). The correlation of z_i with $w_1z_1 + w_2z_2 + \ldots + w_nz_n$ is

$$(11) \qquad R_i = \frac{w_i + \sum_{j=1}^{n} w_j r_{ij}}{\sqrt{\sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij}}} \; . \quad (i \neq j)$$

Setting all such R's equal to some arbitrary constant p and solving for the w's is equivalent to setting the numerators equal to p and solving for the w's since the denominator is a constant. This method is logically defensible only if none of the variables are negatively correlated.

Minimum Generalized Variance. Wilks (1938) has proposed a method of minimizing generalized variance, an analogous extension of the concept of variance, to obtain a set of weights for combining a number of component variables to form a composite when there is no external criterion. In an n-dimensional space the score of a single individual may be represented as a point whose projections on the n coordinate axes correspond to the scores obtained by the individual on each of the n component variables. An n-dimensional simplex may be determined in this space by taking n points plus the point representing the mean of all n variables. The generalized variance is found by squaring the volumes of all such simplexes formed by taking different combinations of n points, summing, taking the mean, and multiplying the result by $(n!)^2$. In the case of one variable, this is the variance, where $(n!)^2 = 1$, and instead of squaring volumes it is the length of the line segments connecting single points with the mean of the distribution which is squared. In the case of two variables it is the area of all possible triangles formed by pairs of points and the point representing the mean of both distributions which is squared and averaged and multiplied by $(2!)^2$. In the case of three dimensions it is the volume of all possible tetrahedra form by trios of points and the point representing the mean of the three variables which is squared.

Briefly, Wilks' method is applied to the weighting problem as follows. An n-dimensional space is defined by the n component variables x_i. The score of the pth person on the ith variable may be denoted x_{ip},where all scores are in standard form. A linear function of the x_i's is sought, such that for any given value of the function the generalized variance of individuals having that value is minimized. The "plane" $T = w_1x_1 + w_2x_2 +\ldots+ w_nx_n$ cuts across the n-dimensional space, determining a series of n-1 dimensional spaces which are non-intersecting. The generalized variance of individuals within each of these subspaces is then minimized and a single set of weights found which satisfies this condition.

Minimum Variation. In 1936 Edgerton & Kolbe presented a method for combining a number of measures of the same thing based on the criterion that the sum of the squares of the $n(n-1)/2$ differences between standard scores for an individual on each of the n variables be a minimum. In other words, intra-individual differences in standard scores are minimized. In the same year a method was suggested by Horst for deriving a set of weights which would maximize the difference between composite scores for all pairs of individuals, *i.e.*, maximize inter-individual differences. Interestingly, this approach leads to weights which are proportional to those obtained via the former criterion. Edgerton & Kolbe, noting that the two methods yield identical results, maintained that their method was computationally simpler.

Maximum Reliability. In the absence of an external criterion, probably no alternative criterion has so frequently been seized upon as maximum reliability. Weighting for reliability has been especially popular when the variables to be weighted are tests which comprise a battery or the items of a single test. It is well known that the maximum correlation which may be obtained between two variables is limited by their respective reliabilities: $\rho_{xy} = \rho_{tt}\sqrt{\rho_{xx}\rho_{yy}}$, where ρ_{tt} is the correlation between "true" scores. In measurement, reliability is the *sine qua non*, the necessary but not sufficient condition, for a valid instrument. For this reason weighting for maximum reliability has long been deemed a worthy enterprise.

When reliability is defined in terms of the proportion of total composite variance which is "true score" variance (or, 1 - the proportion which is error variance), the sample reliability coefficient of the composite y is given by the following formula when all variables are expressed in standard form:

$$(12) \qquad r_{yy'} = \frac{\sum_{i=1}^{n} w_i^2 r_{ii'} + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij}}{\sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij}}, \quad (i \neq j)$$

From this formula it is apparent that the reliability of the composite may equal 1.00 if and only if every $r_{ii'}$ also equals 1.00. Likewise, if every $r_{ii'}$ is zero, $r_{yy'}$ must also be zero. Mosier (1943) has discussed the effect on $r_{yy'}$ of the interrelationships among the variables. For example, if the variables are mutually uncorrelated, the reliability of the composite is the weighted mean of the item reliabilities $r_{ii'}$, where each $r_{ii'}$ is weighted by w_i^2. He notes that this conclusion is of particular interest because when multiple regression is used for prediction, every attempt is usually made to obtain predictors which are independent or nearly so. It may be noted from Equation (12) that for a given set of individual reliabilities and weights, the reliability of the composite increases as the positive intercorrelation of the components increases, although the unreliability of the components does set an upper limit to the size of these correlations.

Equation (12) may be conveniently expressed in terms of two matrices, r and R, both of which contain the component intercorrelations in the off-diagonal cells, and the row vector of weights w_i. In the diagonal cells of r are the reliabilities $r_{ii'}$, whereas in R the diagonal elements are unity. Using this notation Equation (12) becomes

$$(13) \qquad r_{yy'} = wrw'/wRw'.$$

Thomson (1940) derived formulas in matric form for both the maximum battery reliability and the weights which give this result. Peel (1947, 1948) has shown that Thomson's formulation may be considerably simplified. Maximum reliability may be found by solving $|r - \lambda R|$ for its largest root λ_1. The desired weights are then in the ratio of the elements of any row of $\text{adj}(r - \lambda_1 R)$. Peel (1948) has also given equations for the weights which will maximize the correlation between a predictor battery and a complex criterion, itself a weighted composite with fixed weights.

Validity vs. Reliability. Since methods are available for computing both the weights which give maximum validity and the weights which give maximum reliability, an interesting question is, "What is the effect on reliability of weighting for validity, and *vice versa*?" Since the two sets of weights are not at all likely to be proportional, weighting for one criterion, *e.g.*, validity, will result in a less-than-maximal value of the other.

This general lack of correspondence between the two sets of weights may be attributed in the main to two factors. First, *when other factors are held constant*, weighting for validity results in weighting more heavily those variables which are more highly correlated with the criterion. Likewise, weighting for reliability weights more heavily the more reliable variables. Thus, unless the more reliable variables are also the more valid ones (with respect to the *observed* correlation with the criterion) the correlation between the two sets of weights will not be perfect or even nearly so.

Let us consider one case where this is likely to be true and one case where it is not. If all the items of a test are assumed to measure the same thing except for error of measurement, and if all are of a constant level of difficulty, then differences in observed correlations with the criterion are due solely to unreliability, *i.e.*, if all correlations were corrected for attenuation, they would equal a constant, the "true" correlation with the criterion. In this case, the items with the highest reliability will also have the highest observed correlation with the criterion. Thus, as far as this influence on the relative sizes of the weights is concerned, weighting for validity should have the effect of increasing reliability as well and *vice versa.*

However, it is not difficult to conceive of instances where the most reliable items of a test are not the most valid and the most valid not the most reliable. For example, in a multiple-choice test with items of different degrees of difficulty, some items which are very easy may be passed by nearly all examinees. These items may have higher reliability but lower validity than do some of the very difficult items in the test. (Very difficult items tend to have low reliability because of guessing.) Contrasting only these subsets of items from the test, weighting for reliability would weight the easy items higher than the difficult ones, and weighting for validity would do the opposite.

In actual practice it is not likely that many of the items of the test would behave in this manner. This is due partly to the fact that item unreliability prevents extremely high correlations with the criterion, thus making it unlikely that the very unreliable items would have high observed correlations with the criterion.

The second factor which affects the two sets of weights differently is the intercorrelation of the component variables. As noted earlier, when other factors are held constant, the variables which are more independent will receive higher weights in the multiple-regression case. When weights are derived to maximize reliability, however, high positive intercorrelation of the components provides stability and thus, other things equal, the components which have higher correlations with the remaining components are weighted more heavily.

This may be seen readily in the formula which Mosier (1943) derived for the weight to be assigned to the *p*th item, in order to maximize reliability, when the qth item is taken as a reference and assigned a weight of 1.00:

$$(14) \qquad w_p = \frac{\sum_{i=1}^{n} w_i r_{ip} (1 - r_{qq})}{(\sum_{i=1}^{n} w_i r_{iq})(1 - r_{pp}) + r_{qq} + r_{pp}} . \qquad (i \neq p,\ i \neq q)$$

Note that the sum of the intercorrelations of the reference item appears in the denominator of all weights and that the sum of the intercorrelations of the pth item appears in the numerator. Of two items with the same reliability, the one with the higher total intercorrelation with the other items will have the higher weight. Thus two factors, item validity *vs.* item reliability, and total intercorrelation of an item with the remaining items, work against the perfect or near-perfect correlation of the sets of weights which maximize validity and reliability respectively.

Table 1

The Effects of Weighting on Validity and Reliability

Correlations	I	II	III	IV
r_{11}	.50	.40	.50	.50
r_{22}	.95	.60	.95	.95
r_{01}	.20	.20	.40	.20
r_{02}	.40	.40	.20	.20
r_{12}	.10	.10	.10	.10
Unweighted				
Validity	.405	.405	.405	.270
Reliability	.660	.545	.660	.660
Weighted for Validity				
w_1	.162	.162	.384	.182
w_2	.384	.384	.162	.182
Validity	.460	.460	.460	.270
Reliability	.884	.598	.597	.660
Weighted for Reliability				
w_1	.010	.154	.010	.010
w_2	1.000	1.000	1.000	1.000
Validity	.402	.425	.204	.202
Reliability	.950	.607	.950	.950

In Table 1 we present numerical illustrations of some of the foregoing points. Four hypothetical sets of data, including the reliabilities, validities, and intercorrelation of two variables which comprise a two-test battery, appear in the first section of the table. Below these, the validity and reliability of the unweighted composite are given, followed by the weights and resulting validities and reliabilties obtained when the composite is weighted for validity and reliability respectively.

In Case I, the reliability and validity are in the same direction for the two tests, *i.e.*, the more reliable test has the higher observed correlation with the criterion. Thus, it is to be expected that weighting for either validity or reliability will increase both. But this is *not* the case. Weighting for validity does increase reliability from .660 to .884, but weighting for reliability *reduces* the validity, despite the fact that the more reliable test is also the more valid. The reason for this is apparent from the sizes of the weights. Since the second test is considerably more reliable than the first, and since the intercorrelation of the tests is low, the weight assigned to the second test is 100 times as large as that assigned to the first, thus all but eliminating the first test from the composite, despite its small independent validity. In the unweighted case, the second test by itself correlates .400 with the criterion, and adding the first test to the battery increases this correlation very slightly to .405. Weighting for reliability waters down this contribution greatly, resulting in a gain of only .002 over the correlation for the second test alone. The reliability of this composite, however, is virtually identical to the reliability of the more reliable test.

In Case II the general pattern of the correlations remains the same, but here the difference between the reliabilities of the tests is not so great. In this case the expected trend does occur. Weighting for validity produces the same regression weights and the same resultant validity as in Case I, although the increase in reliability is smaller because of the reduced reliability of the individual tests. Weighting for reliability, however, produces weights which are in the ratio of approximately 13:2 rather than 100:1, again increasing the reliability of the composite to a value almost as high as the reliability of the more reliable test. The validity is increased modestly from .405 to .425.

Case III illustrates the situation where the more reliable test has a very low correlation with the criterion, whereas the less reliable test, despite its unreliability, has a higher correlation with the criterion. In this case, as expected, weighting for validity reduces the reliability from that of the unweighted composite. Likewise, weighting for reliability reduces the validity from that of the unweighted composite. Again, since the two tests differ so greatly in reliability, weighting for maximum reliability reduces the contribution of the less reliable test to the validity of the composite to nearly zero.

Case IV illustrates the fact that when tests which differ in reliability but have equal correlations with the criterion are weighed for reliability, the validity of the composite is reduced from that computed for the unweighted composite. In this case, since the observed correlations with the criterion are equal, weighting for validity produces equal regression weights and no increase in the validity of the composite. In this case, *any* weighting scheme other than equal weighting of the two tests will produce a reduced validity for the composite. Of course if there were more than two variables in the composite, differences in their intercorrelations with one another *would* produce differences in the regression coefficients and some increase in validity.

The above illustrations are at best oversimplified because it is not too often that one combines only two measures. In any situation where there are more than two variables.to be combined, the individual validities, reliabilities, and intercorrelations will interact to produce the resultant effect on one measure of weighting for a maximum value of the other.

From Table 1 it is clear that weighting for reliability cannot increase the reliability of the composite to a value higher than the reliability of the most reliable test in the battery. What weighting actually accomplishes is suppression of the contributions of the less reliable variables, leaving it to the most reliable tests to constitute the composite score. If, on the other hand, the less reliable tests could, by some means, be made more reliable, the reliability of the unweighted composite would automatically rise, as would the validity.

Table 2

The Effects of Weighting for Validity with Perfectly Reliable Tests

Correlations	I	II	III	IV
r_{01}	.283	.316	.566	.283
r_{02}	.410	.516	.205	.205
r_{12}	.145	.204	.145	.145
Unweighted				
Validity	.458	.537	.510	.322
Reliability	1.000	1.000	1.000	1.000
Weighted for Validity				
w_1	.228	.220	.548	.259
w_2	.377	.472	.126	.168
Validity	.468	.552	.588	.333
Reliability	1.000	1.000	1.000	1.000

In Table 2 the correlations of Table 1 have been corrected for attenuation. These values are thus the observed correlations which would be expected if all the tests were perfectly reliable. With perfectly reliable predictors the composite would also be perfectly reliable and the unweighted validity of the composite would increase accordingly. These perfectly reliable tests might then be weighted for maximum validity. The resulting validity coefficient would be the maximum corelation obtainable with these tests.

Since actually making the tests more reliable automatically increases the validity, whereas weighting for reliability may or may not increase validity, it would seem that it is always safest to attempt to increase reliability *per se*, rather than to weight for increased reliability. This would definitely be preferable if a criterion measure is available. If no criterion measure is available, and hence the validity is unknown, weighting for reliability is to be recommended only if it may be safely assumed that the most reliable tests are not, in fact, the least valid ones.

It may be objected, however, that since reliability is a necessary prerequisite for a valid test, then if there is no criterion measure, increasing the reliability of the test is always to be desired. Somewhat similarly, it has been suggested that, even when criterion measures are in hand, one might wish to increase *both* validity and reliability by stipulating that these shall be equal and then solving for the weights which maximize this value (Thomson, 1940). Both of these positions seem to advocate increasing reliability even at the cost of some validity. Admittedly, the unreliability of a test sets an upper limit to the validity of the test. But if a test with low reliability correlates more highly with the criterion than a quite reliable one does, then despite its unreliability the test will always be expected to correlate more highly with the criterion. Likewise, an unweighted composite of these tests will always be expected to correlate more highly with the criterion than a composite weighted for reliability. What the unreliability of the more valid test *does* do is prevent the maximum correlation for this test with the criterion from occurring. But manipulating the reliability of the test statistically is not necessarily a good thing, particularly since validity may be lowered in the process. These statements apply even when the actual validity is not known, thus explaining why weighting for reliability exclusively is not recommended unless it may be assumed on other grounds that this will not decrease the real but unknown validity.

In testing, most measures, even if highly reliable, do not often have extremely high correlations with the criterion. Thus, it is unlikely that we will discover many variables which, while unreliable, are nevertheless more valid than their more reliable companions in the composite. In the usual situation a reasonable reliability is needed before the test or test item can evidence *any* validity at all. It is probably for this reason that reliability and increasing reliability have received so much attention. But it must be emphasized that once the validity is known, reliability must assume a position of secondary importance. It is better to have a test with reliability of .60 and validity of .57 than a test with reliability of .95 and validity of .19. In the former instance the "true" correlation with the criterion is .95, whereas in the latter it is .20!

These, then, are some of the more important methods which have been used to assign weights to the component variables which comprise a composite. A number of additional weighting methods deserve mention. Some of these are admittedly approximations to multiple regression weights and others are simply weights which have been used for one reason or another. Each will be discussed briefly before we move on to the next major section of the paper.

Weighting by the Reciprocal of the Standard Deviation. Quite frequently the authors of tests wish to eliminate the influence of unequal standard deviations on the effective weighting of a number of variables. Weighting each measure by the reciprocal of its standard deviation accomplishes this. Using standard scores has the same effect and in addition subtracts out the mean from each measure. Some testers have mistakenly believed that this ensures equal weighting. This conclusion is unwarranted, of course, unless the component variables are uncorrelated or equally correlated. Otherwise the intercorrelations will determine the effective weights.

If no particular significance is attached to the fact that the several variances are unequal, then removing this source of unintended weighting may be appropriate. There is at least one case, however, where this would not be true. Richardson (1941) presents an example similar to the following. Suppose X_1 is the number of items answered correctly on a 50-item test and X_2 is the number of items answered correctly on a 100-item test. X_2 will undoubtedly have a larger variance than X_1. But it is also true that the longer test will in general be a more reliable test. If the scores are merely combined, the longer test will automatically have the larger effective weight. This will work in favor of the reliability of the composite. In this case, weighting by the reciprocal of the standard deviation denies any such difference between the tests and thus works against the reliability of the composite. If the two tests in the composite measure the same thing, then the increased reliability of the longer test would also be reflected in a larger validity, again arguing against the use of these weights.

Weighting by Length. The above example raises the question of whether or not tests should be weighted in terms of their length. Originally, the idea of weighting by length can probably be traced to the fact that examination grades are often expressed as percentages of items answered correctly. Combining such percentages directly gives equal nominal weighting to each test. But clearly, if one test consists of 50 items and a second of 100 items in the same subject, then the second test is, in a very real sense, equal to two of the first. Weighting each percentage in terms of the length of the test on which it was computed has the effect of converting the percentages back to a score equal to the number of items answered correctly. By so doing each item now has equal nominal weighting[3]. By simply adding the percentages the tests will be equally weighted, but the items will not.

But what of the case where the tests are not in the same subject, but instead tests from a number of different subjects are to be combined to form some sort of overall achievement score? In this case it is not so clear that each item should be equally weighted. The principal reason for this is that the significance of a single item may differ markedly from one subject to the next. A single lengthy algebra problem simply cannot be considered equivalent to a single vocabulary item. The item is a meaningful unit only when the items are measuring the same thing or very similar things.[4] In such a case it would undoubtedly be better to work with the percentage scores, perhaps weighting these on the basis of other *a priori* or empirical considerations.

Weighting by Difficulty. Another method of weighting which has been popular, particular in the classroom, is weighting by difficulty. Very often such weighting is implied rather than explicit, as when a teacher assigns different weights to sections or items of a test on the basis of an intuitive feel for the difficulty or "worth" of the component in question, rather than some conviction concerning the intrinsic validity of the component. In other cases, particularly with some standardized tests, the weights are derived via an empirical estimate of the difficulty of the item. In these cases the weight is usually equal to the proportion of those taking the test who fail to answer the item correctly.

The logic of this type of weighting is most likely based on the con-

[3] See footnote 1.

[4] Likewise, for items which do measure the same thing, it might also be argued that items of different forms, *e.g.*, true-false *vs.* multiple-choice, are not equivalent units.

viction that knowing a very difficult item is evidence of considerably more ability or achievement than knowing a simple one. But no one seems to have pointed out that, in effect, this is the same as penalizing the student more heavily for not knowing a difficult item than for not knowing an easy one, a rather counter-intuitive strategy. If the weights were reversed we would be in the position of penalizing the student more for missing an easy item than a difficult one, but at the same time allowing less credit for a correct answer to a difficult item than an easy one. As long as there is but a single set of weights which is monotonically related to difficulty, we cannot have one side of the coin without the other. One possible way around this difficulty, however, would be to give more credit for passing a difficult item than an easy one, and at the same time to penalize more severely, with a negative weight, for missing an easy item than for missing a difficult one. For example, let the positive weight equal q, the proportion of examinees failing the item, and let the negative weight equal $-p$, where p is the proportion passing the item. Thus a difficult item passed by only .05 of the examinees would be scored .95 if passed and -.05 if failed. The mean score for each item over all examinees is $qp + (-pq) \equiv 0$ and thus the mean test score for all examinees is also zero, although the distribution of the scores will depend on the distribution of the item difficulties. This scheme does not, of course, take guessing into account, except insofar as the values of p and q are affected by guessing. Although this weighting scheme is *not* being recommended, it is logically more defensible than simply assigning weights according to difficulty.

It is interesting to note that when items varying in difficulty are given equal nominal weights a certain amount of natural weighting-by-difficulty occurs, although this weighting is nct a monotonic function of difficulty. As the difficulty of an item deviates from .50 the maximum phi coefficients for that item with the other items of the test becomes smaller, thus limiting the size of its maximum possible contribution to total variance. Thus the most and least difficult items tend to be less heavily weighted than items of .50 difficulty.

Weighting by Validity. When it is not feasible to carry out a full-scale multiple regression derivation of appropriate weights for component variables, very often the raw correlation of the component with the criterion is used as an approximation to the optimal weight. Such a weight ignores the intercorrelation of the components and the variance of the individual component being weighted. If, however, standard scores are weighted in this manner, the intercorrelation of the components is the only factor which is left unaccounted for. Since in a single test the items are usually fairly homogeneous and the average intercorrelation of any one item with the others may be fairly constant (particularly if items are of similar difficulty), the approximation may be a very good one indeed. The same is true if the components are nearly independent of one another. These weights are in least correspondence with the multiple regression weights when the average intercorrelation varies markedly from one component to the next and when raw scores are used which differ markedly in variance.

Guilford (1941) has presented a formula for weighting test items which is an approximate regression weight for X_i and which combines the correlation of X_i with the criterion c, the standard deviation of the criterion, and the standard deviation of X_i:

$$(15) \qquad w_i = r_{ic}s_c/s_i.$$

The only factor not included in this weight is the intercorrelation of the items. Guilford goes on to simplify this expression by assuming criterion groups of equal size, thereby fixing s_c at .50. The formula, after simplifying and transforming to achieve a range of weights from 0 to 8, is

$$(16) \qquad w_i = (P_u - P_l)/P(1 - P) + 4,$$

where P_u is the proportion of people in the "upper" criterion group who select the response and P_l is the proportion in the "lower" group who select the response. P is the proportion in the combined group $P_u + P_l$ who choose the response. As Guilford presented the method it was intended for use with responses, but as can easily be seen, it also lends itself readily to use with dichotomously scored items.

At least one other index of validity has been used to weight the items of a test. Clark (1928) presented a formula for *evaluating* the items of a test. His index of validity, IV, was given by

$$(17) \qquad IV = (P - D)/(1 - D),$$

where D is the percentage of the group taking the test who fail the item and P is the percentage of the "criterion group" who fail the item. For a given item the criterion group is composed of the D percentage of the class who rank lowest in terms of total score. Although Clark seems to have intended his Index of Validity as a measure of the "goodness" of a particular item, at least one person, Peatman (1930), has used it to weight items.

Factor Analysis. One further method of weighting deserves mention. If no criterion measure is available, a correlation matrix may be factor analyzed to extract the major factors accounting for the variance of the unweighted composite. Factor scores, correlated or uncorrelated, may then be secured. See Glass and Maguire (1966) and Harris (1967).

The Effectiveness of Weighting

In each situation where a set of weights is used with a set of variables the specific effect of using one particular set of weights as opposed to another is uniquely determined by the factors which have been discussed under

"*Methods of Weighting*." However, it is possible to make certain generalizations concerning the *limits* of the effectiveness of any set of weights relative to another set of weights regardless of the method used to derive either.

It is well known that if the correlation of each of two variables with a third is known, then the limits of possible values of the correlation between the two variables is determined.[5] For example, if each of two variables correlates .90 with a third variable, then the correlation between the two variables must lie within the range $.62 \leq r \leq 1.00$. Therefore, if the correlation between one weighted composite and a criterion is known, and if the correlation between the two weighted composites is known, then the limits of the correlation of the second composite with the criterion is determined. However, even in the absence of information concerning the size of the correlation of one composite with the criterion, the size of the correlation between the two composites gives some indication of the limits of the effectiveness of either weighting method over the other. If it is known that two different sets of weights produce composites which correlate .99, then regardless of the correlation of either composite with the criterion, adopting the alternative set of weights could not be expected to affect that correlation very greatly.

A number of authors, notably Wilks (1938), Richardson (1941), Burt (1950), and Gulliksen (1950), have presented formulas for the correlation of two weighted sums. Since Gulliksen's formula is the most general, and since he discusses important special cases, it is closely followed here. If n standard scores are weighted by the weights v and the same set of scores also weighted by the set of weights w, we may denote the respective composite scores X_v and X_w. Without yet imposing any restrictions on the weights, the correlation between the two composites X_v and X_w is given directly by the following formula:

[5] See Stanley & Wang (1969)

$$(18)\quad r_{X_vX_w} = \frac{\sum_{i=1}^{n} v_i w_i + \sum_{i=1}^{n}\sum_{j=1}^{n} v_i w_j r_{ij}}{\sqrt{\sum_{i=1}^{n} v_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} v_i v_j r_{ij}}\sqrt{\sum_{i=1}^{n} w_i^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j r_{ij}}} \quad (i \neq j) \; .$$

In Equation (18) the weights have been expressed in raw-score form. The sums of squares and cross-product sums may, however, be expressed in terms of the means, variances, and covariances to which they are related. When this is done and the expression is simplified, Equation (18) becomes

$$(19)\quad r_{X_vX_w} = \frac{n(1-\overline{r}_{ij})(\mathrm{Cov}(v_iw_i) + \overline{v}\overline{w}) + (n^2-n)\mathrm{Cov}((v_iw_j)r_{ij}) + n^2\overline{v}\overline{w}\overline{r}_{ij}}{\sqrt{\begin{array}{l} n(1-\overline{r}_{ij})(\sigma_v^2+\overline{v}^2) \\ + (n^2-n)\mathrm{Cov}((v_iv_j)r_{ij}) \\ + n^2\overline{v}^2\overline{r}_{ij}\end{array}}\sqrt{\begin{array}{l} n(1-\overline{r}_{ij})(\sigma_w^2+\overline{w}^2) \\ + (n^2-n)\mathrm{Cov}((w_iw_j)r_{ij}) \\ + n^2\overline{w}^2\overline{r}_{ij}\end{array}}} \; ,$$

where, as usual, $i \neq j$. From this equation it may be seen that the correlation between the two weighted composites depends upon the number of scores to be combined, n; the mean values of the two sets of weights, $\overline{v}$ and $\overline{w}$; the variance of the two sets of weights, σ_v^2 and σ_w^2; the average intercorrelation of the variables to be combined, $\overline{r}_{ij}$; the covariance between the two sets of weights, $\mathrm{Cov}(v_iw_i)$, and the covariance of a product of weights with a corresponding correlation, $\mathrm{Cov}((v_iw_j)r_{ij})$. To see what happens to this expression as n increases we may divide the numerator and denominator by n^2 and eliminate all terms which have $1/n$ as a factor. Thus,

$$(20)\quad r_{X_vX_w} \Rightarrow \frac{\mathrm{Cov}((v_iw_j)r_{ij}) + \overline{v}\overline{w}\overline{r}_{ij}}{\sqrt{\mathrm{Cov}((v_iv_j)r_{ij}) + \overline{v}^2\overline{r}_{ij}}\sqrt{\mathrm{Cov}((w_iw_j)r_{ij}) + \overline{w}^2\overline{r}_{ij}}} \; . \quad (i \neq j)$$

This expression will be equal to unity if the covariance terms are equal and if the mean v weight equals the mean w weight. If the covariance terms are nearly zero, such that they may be ignored, then the correlation approaches unity regardless of the mean value of the weights. The information concerning Equations (19) and (20) has been summarized by Gulliksen as follows:

1. If either or both $\bar{v}$ and $\bar{w}$ may be zero, $r_{X_v X_w}$ may assume any value regardless of the value of $\bar{r}_{ij}$, n, or the various covariance terms involving the weights.

2. If v and w are small in relation to σ_v and σ_w, $r_{X_v X_w}$ depends primarily on the four covariance terms and is relatively insensitive to changes in the values of $\bar{r}_{ij}$ and n.

3. If we consider only positive weights so that $\sigma_v/\bar{v}$ and $\sigma_w/\bar{w}$ are less than unity, the correlation between the two composites obtained by using two different sets of weights approaches unity as (a) the correlation between the two sets of weights is increased, (b) the average intercorrelation of the component variables is increased, and (c) the number of component variables to be combined is increased. It should be particularly noted that the last effect holds, even if the correlation between the two sets of weights is zero, provided $\bar{r}_{ij}$ is greater than zero. (d) As the standard deviation of the weights is increased in proportion to the mean weights, $r_{X_v X_w}$ approaches unity regardless of the values of $\bar{r}_{ij}$, $\bar{v}$, and $\bar{w}$.

From these deductions it is clear that there are very real limits on the effectiveness of any weighting method, particularly when the number of predictor variables is large and only positive weights are used. Under these conditions even random sets of positive weights will result in composites which are highly correlated. When the weights have been derived according to some logical rationale, the correlation is likely to be very high indeed.

Gulliksen concludes that from a practical point of view, 50-100 variables is probably sufficient to make differential weighting unprofitable, and the same conclusion is reached if the variables are very highly correlated. Weighting may be worthwhile, he contends, when there are few, say three to ten, variables to be combined and if the average intercorrelation is also low, say .50 or less. However, in addition, even in this case, the weights must have an appreciable standard deviation if they are to differ from unit weights appreciably. And finally, if two sets of weights are being considered, and the weights themselves are highly correlated, it will make little difference which set is used.

A word of caution is in order concerning the wholesale dismissal of the weighting question under conditions of high correlation between differently weighted composites. It was pointed out earlier that the *limits* of the effectiveness of one set of weights given the effectiveness of another set and the correlation between the two weighted composites is easily determined. It is implicit in the correlation-between-weighted-composites approach that if the correlation rapidly approaches unity, then it really doesn't matter which set of weights is used. This is only partially true. As the validity of one weighted composite drops from unity, the range of possible values which another weighted composite may give when correlated with the criterion increases, with a constant correlation between the composites. Likewise, for a constant validity of the first weighted composite, as the correlation between the two weighted composites drops from unity, the range of possible validities for the second increases. McCornack (1956) has criticized a great many empirical studies of the effectiveness of weighting for failure to take this into account. Thus, quite often investigators are content to report only that two composites correlate over .90 or some higher figure without reporting or even investigating whether one composite is more or less valid than the other. Yet this is the result that is of most importance. In such cases, and they are numerous, conclusions to the effect that weighting is not worth the trouble or that two methods of weighting will result in essentially equivalent scores, while probably correct, are not completely justified.

A word of caution is also in order concerning a very similar premise, *viz.*, that if the correlation between two composites is greater than the reliability of either, then it does not really matter which composite is used. The argument which has been advanced by many, including Burt (1950), is that if two tests (composites) are more highly correlated with one another than either is with itself on another administration, then either one should be acceptable. But this still does not allow for the real possibility that one version will have a higher validity than the other. This will probably not be the case, but it nevertheless must be recognized as a possibility.

So far a number of factors which influence the effectiveness of weighting have been considered: the number of variables in the composite, their average intercorrelation, the size of the weights and their correlations, *etc.* Nothing has been said, however, concerning the size of the sample on which the weights are derived or the distribution of the scores on the several component measures. This section will be concluded with a brief mention of these two points.

In the section on multiple regression it was pointed out that when weights are derived on the basis of a sample from the population of interest, sampling error will cause a certain amount of shrinkage in the validity coefficient when the weighted composite is used to predict the criterion in the population. It is usually recommended that if multiple regression weights or similar weights are to be empirically derived and used on a wide scale, the sample on which the weights are derived should be fairly large. In deriving the weights for responses to the Strong Vocational Interest Blank, Strong recommended, for example, that no less than 250 blanks for each occupation should be used (Strong, 1943). In many personnel situations, and even in the classroom, numbers this large are usually out of the question. However, the lack of stability which small sample size brings about may also be responsible for the observed lack of difference between different weighting methods.[6] In an interesting empirical study of the effects of sample size on the predictive validity of the resulting composite, Lawshe & Schucker (1959) found no difference between samples of 20, 40, and 90 cases when the weights were used to predict the criterion in a cross-validation sample. They concluded, however, that more research on sample size is needed.

Although we have not stated it explicitly until now, most of the weighting methods in common use do assume that the measures on the several component variables are normally distributed, or at least have similar distributions. Such assumptions are most important when tests of significance are performed or when point estimation is involved. Failure to satisfy an assumption of normality may have other consequences. For example, Cliff (1960) investigated the effect of unlike distributions on the contribution to composite variance made by two tests which formed a composite. One test was negatively skewed and the other was positively skewed, although the tests had the same variance since standard scores were used. It was hypothesized that summing the standard scores to get a composite would not result in equal contributions to composite variance at various cutting points in the composite score distribution. By actually computing the contribution of each variable to the composite variance it was demonstrated that the positively skewed variable contributed more to composite variance in the upper percentiles and the negatively skewed variable contributed more in the lower percentiles, whereas if the variables

[6] R.G. Simpson (1951) has considered the sample problem at length with reference to the weighting of biographical inventory items.

had been symetrically distributed, they would have contributed equally throughout the distribution.

Methods of Weighting Response Categories

In each of the weighting methods discussed thus far the entity which was weighted, X_i, was a quantitative variable capable of taking at least two values. Each subject received a score on each of the n component variables and it was these scores which were then weighted to determine the composite score. In the present section we will consider the case where for each item X_i we can categorize the subject's response into one of a small number of mutually exclusive response categories which do not initially have numerical values associated with them. The weighting problem is one of determining a set of weights for the categories in order to derive a total score for the subject. Conceptually, the problem is not very different from that of scaling the response categories in order to assign to a subject the scale value of the category he selects.

The response-weighting methods to be discussed may be classified according to the nature of the criterion which is used to derive the weights. We will first consider methods which utilize an external criterion which is classificatory. Next we will consider the use of an external criterion which is quantitative, and finally we will turn to the use of an internal quantitative criterion.

Weighting with an External Qualitative Criterion. Consider a single stimulus, X_i, to which a subject's response may be classified in one of c mutually exclusive categories. The stimuli might be personal, biographical, or demographic questions, the items of an interest or personality test, or any such similar thing. As criterion measures we have the responses of two or more criterion groups to the stimulus and we wish to determine weights for each of the *c* categories in order to best estimate whether the subject's response is more typical of one criterion group or the other. Although *a priori* weighting of responses is occasionally found (*e.g.*,see Giles,1936), most often the weights are derived empirically. In one case, Gage (1957) found that empirically derived response weights were not superior to logically assigned weights with respect to the reliability and validity of the composite test. Nevertheless, in many cases it is not possible to determine logically which responses are to be weighted most heavily. Typical of these situations is the interest test.

Strong (1943) has presented an historical survey of methods of weighting responses of an interest test. His discussion is the basis of the brief summary of these methods which follows. These methods have in common the fact that the criterion to be predicted is qualitative, usually membership in a particular group. Although Strong was concerned specifically with the responses "dislike," "indifferent," and "like" (which could, if desired, be ordered, *e.g.*,0,1,2), the methods themselves are appropriate whenever a number of mutually exclusive categories of response are weighted for diagnostic purposes.

In 1924, Ream used the following rationale to weight the *items* of an interest test. He had a successful group of life insurance salesmen and an unsuccessful group respond "like" "dislike," and indifferent" to a series of items. He then calculated for each response to each item the proportion of those in each of the two reference groups who selected the response. Whenever the difference between two of these percentages for one of the responses to an item exceeded the standard error of the difference, the item was retained and the score assigned to it was +1 if the direction of the difference favored the successful group and -1 if the reverse was true. This was equivalent to setting the critical ratio equal to 1.00 and weighting zero all items which failed to attain this. However, Ream did not weight the several responses to a single item differentially.

A much more recent example of a very similar approach to response weighting is found in Anastasi, Meade & Schneiders (1960). In this case the response weights were determined according to the significance of the difference between the proportions of those in the reference groups choosing the response and the direction of that difference.

A somewhat different method was used by both Cowdery (1925) and Strong (1930), based on a formula for weighting recommended by T.L. Kelley:

$$(21) \qquad w_r = r_{rc}/(1 - r^2_{rc})s_r,$$

where r_{rc} is the correlation between choosing the response in question and being in the criterion group, and s_r is the standard deviation of the response distribution. The r_{rc}/s_r part of the formula is actually an approximation to the multiple regression weight which would be assigned to the response. $(1 - r^2_{rc})$ is proportional to the square of the standard error of r_{rc}. In practice the weight is usually multiplied by 10 to get rid of decimals and then taken to the nearest integer. Both Cowdery and Strong used the formula, although procedural differences in presenting the data resulted in different working formulas.

In 1934 Kelley revised the formula, stating that instead of being proportional to the square of the standard error of r_{rc}, the multiplicative constant should be proportional to the square of the error of the weight itself, r_{rc}/s_r. An appropriate formula was derived and the new formula was adopted by Strong for scoring the Strong Vocational Interest Blank. The whole notion of incorporating such a constant in the weight was subsequently criticized by Guilford (1941), however. He claimed that the reliability of the regression weight should have nothing to do with the size of the contribution of the variable to the score. His argument bears a similarity to that advanced in the last section against weighting for reliability.

Strong and, somewhat later, Kuder(1934) used the contrast of the group in question with the composite of other groups, thus having a dichotomous criterion. Porter (1965) tried a contingency-table chi weighting procedure for securing weights from several criterion groups (such as foresters, clinical psychologists, social workers, dentists, and pharmacists) simultaneously to test the hypothesis that for similar occupations his procedure would differentiate better than Kuder's, whereas for dissimilar occupations it would not. Though somewhat equivocal because of an incorrect key to one of the interest scales, his findings tended to support this hypothesis.

He considered Kuder Preference Record -- Vocational items that required three things to be ranked by picking the one liked most and the one liked least. For example, an item might consist of three options "Construct a piano," "Play a piano," and "Move a piano." There are six possible orderings of those three phrases, each of which orders can be considered a "response." If there are five different occupational groups, this yields a 6 x 5 set of tallies. Porter simply computed chi (the signed difference between the number of responses in one of the 30 cells and the theoretical number for that cell determined from the respective row and column sums, divided by the square root of the theoretical number).

These figures, which resemble percentage deviations from expectancy, were his option weights. A given examinee would obtain five scores for his pattern on a single item, *i.e.*, one score for each of the five occupational groups. Porter used every item in the Kuder Preference Record, merely summing an examinee's pattern weights for each occupation to yield a score scale for that occupation.

Weighting with an External Quantitative Criterion. In 1941 Louis Guttman discussed at length the weighting of response categories. He showed that if we wish to predict a quantitative external criterion y by assigning weights to each of a number of response categories x_c, the correlation ratio $\eta^2_{y\bar{x}}$ will be a maximum if each category is weighted by the mean criterion score of persons in that response category. Such a weighting scheme produces a perfect regression of criterion scores on category scores. The weighting scheme also maximizes the correlation r_{xy}. If a number of items are available, response weights for each may be determined by the above procedure. In order to maximize prediction of the criterion using all of these items, it would then be appropriate to combine them in a multiple-regression equation, if it is assumed that the regression of criterion scores on these item scores is linear. Guttman, however, makes the simplifying assumption that the items are independent and uses the unweighted mean of the category weights to determine the total score for each subject. This procedure does not, of course, take into account the different "validities" of the items or differences in their intercorrelations.

It is interesting to note that although Guttman does not discuss the possibility, there is an alternative approach to the multi-item situation. Recall that in the single-item case the response of the subject was straightforwardly categorized. If there are k permitted responses to an item, the subject's choice determines which of k categories he falls into. Once categorized, the determination of the weights is simple. For n items, each with k alternative responses, it is possible to categorize each subject uniquely by the particular *combination* of responses he selects over the n items. There are k^n such categories possible, and after all subjects have been categorized, the determination of the weights is the same as in the previous case.

It is not particularly surprising that this method was not explicitly suggested by Guttman. As the number of items increases the number of possible response combinations increases exponentially. For 10 five-option items it is 9,765,625! This is a far less economical method than simply using the *name* of the subject to predict his score! Still, for a very few items and a very large number of subjects, the scheme does have the advantage of maximizing prediction.

Wherry (1944) has discussed a special case of the one considered by Guttman. Where the external criterion is expressed as a pass-fail dichotomy, scored 1,0, Wherry shows that for a single item the response weights which maximize the point biserial r are weights which are proportional to the proportion of passers in each response category. This proportion is exactly equal to the mean of the subjects choosing the response, where the criterion scores have been scored dichotomously. For a theoretical study of Guttman's procedure as applied to option-preference patterns, see Merwin (1959).

Weighting with an Internal Quantitative Criterion. Where there is no external criterion with which to weight item responses, it is nevertheless possible to derive weights which will differentiate among subjects with respect to a composite score. Guttman (1941) considers the total scores to have meaning only , insofar as they enable us to differentiate between the candidates consistently. Guttman seeks to maximize the internal consistency of a set of responses to n items. He seeks weights for the response categories which will maximize the correlation over items and subjects between response weight and total score. Clearly, the internal consistency of a set of responses is enhanced if persons with similar total scores tend to endorse similar response categories. Guttman shows that such a correlation is maximized when a score for a person is the mean of the response categories which characterize him, and when the weight for a response category is the mean score of the persons choosing the category.

Clearly, since there is no *a priori* "correct" response, many sets of weights will do. Both Lawshe & Harris (1958) and Shiba (1965) have presented iterative procedures for this case. In the Lawshe & Harris procedure the responses are first given *a priori* weights and each subject's score is calculated by averaging the weights assigned to the responses he has chosen. The weight for each response is then recalculated according to the mean score of those choosing it. The subject's score is then revised according to the new weights and so on until the weights and scores stabilize.

Although there has been no attempt to do so to date, the above response weighting methods, with and without an external criterion, could be used to weight responses to test items where there *is* an *a priori* correct answer. It is generally recognized, at least in theory, that differential weighting of distracters may provide information which is lost when test items are scored dichotomously or with a correction for guessing. If an external criterion were available it would be possible to assign each response option a weight equal to the mean criterion score of individuals choosing the option. Of course, it would be necessary to insure that the correct option for each item has a significantly higher weight than any other option. In the absence of an external criterion an internal criterion such as number of items correct or total score corrected for guessing could be used to weight the options. As in the Lawshe & Harris procedure,it would be possible to continue iterations until scores and weights stabilized. In the former case validity of the items would be maximized, whereas in the latter, reliability would be maximized.

Cross-Validation and Response Weighting. Just as it was true in the case of multiple regression, it is true with optimum response-weighting techniques that weights derived in a particular sample or population have more effectiveness in that same sample or population than any other set of weights. Once again, depending on what the weights are intended to maximize, there is likely to be shrinkage when the original weights are applied to a new sample from a given population. It is perhaps with this in mind that some investigators look for significance before assigning weights. According to Guttman, Strong could simply have weighted response options by the proportion of persons choosing the option who were members of the occupation in question. What Strong did, however, was to consider the *difference* between the proportions choosing and not choosing the option with respect to the profession in question. Since small differences, though real in the sample, might well be due to sampling fluctuations and disappear in a different sample, Strong chose instead to weight more strongly those responses which were more differentiating. The use of the test of significance to determine the size of the response weights in the Anastasi *et al.* (1960) paper is another example of this. Thus, if Guttman's method were to be used in any large-scale testing program, crossvalidation of weights would be extremely important.

This concludes what might be called the analytical approach to the weighting problem. We have seen that alternative definitions of weighting appear in the literature and we have reviewed all of the major methods of weighting which have been and continue to be used. Finally, we have considered, from a rational standpoint, those factors which operate in each concrete situation to determine the effectiveness of the various weighting methods. In the next section we turn to the empirical studies of weighting, those where a specific set of component variables is to form a composite and the problem of whether or not to weight, and/or what set of weights to use, is investigated.

Empirical Studies of Weighting

Empirical studies of weighting far outnumber analytical ones. A great many early testmakers either incorporated weights into their tests as a matter of course (*e.g.*, see Yerkes, Bridges & Hardwick, 1915; Pintner, 1920; Wright,

1929) or tried out one or two methods before making a decision on the weighting question (*e.g.*, see Anderson, 1925; Bovee, Holzinger, & Morrison, 1925). Both the Yerkes-Bridges Point Scale and the Kuhlman-Anderson Intelligence Test incorporated weights of some form. Besides these, a great many less well-known tests incorporated some type of weighting scheme. Because the number of studies is so large, and since the findings tend so strongly in the same direction from one study to the next, each of the following sections of this discussion will be arbitrarily selective. The studies which are to be mentioned in some detail are quite typical of those to be found in the literature.

Although the empirical studies of weighting deal primarily with the weighting of tests, subtests, test items, item responses, and so on, the weighting question has also been explored in other areas. Other types of information to which weighting methods may profitably be applied include economic, anthropometric, and psychological indices (Scates & Fauntleroy, 1938; Stromgren, 1946); biographical or personal inventories (Congdon, 1941; Wherry, 1944); and especially ratings (Bingham, 1932; Jurgensen, 1955; and Tiffin & Musser, 1942). For the sake of simplicity, however, this section will deal only with the weighting of measures which may propoerly be termed "scores" of one type or another. We will first consider the weighting of scores which are themselves composites, *i.e.*, course grades, test scores from a number of different tests or from subsections of a single long test, or test scores from a number of tests of the same thing. Second, we will consider the weighting of the items of a single test, where in most cases the raw score from a single item is in the form of a pass-fail dichotomy. Finally, attention will be turned to the problem of response weighting, both in the interest and personality type of test, where there is no "correct" response, and in the academic-achievement or aptitude type of test, where incorrect responses may be differentially weighted.

The Weighting of Tests

In 1931 Scates & Noffsinger reported a study of factors which influenced the effectiveness of weighting when a number of tests in a battery were to be combined to form a composite. A six-test battery and a ten-test battery were involved. The first battery of tests was given to 80 subjects and the second to 26 subjects. Four methods of weighting were compared: (1) natural weighting using the raw scores on each test; (2) *a priori* weighting based on the opinion of a committee of judges; (3) modified *a priori* weighting; and (4) sigma weighting, *i.e.* weighting by the inverse of the standard deviation. The results were presented in terms of the correlation between composite battery scores under the different types of weighting. For the ten-test battery these correlations ranged from .943 to .985. These correlations are interpreted as evidence against the effectiveness of artificial weighting over natural-raw-score weighting. A fairly high intercorrelation of the tests may explain the high correlation between the composites. As noted above, however, correlation between composites does not deny the possibility of differential validity of the various composites.

A more recent study by Wesman & Bennett (1959) illustrates the more direct approach where the validities themselves are compared for one weighting scheme *vs.* another scheme *vs.* no weighting. In this case the tests were actually the subtests of the Psychological Corporation's College Qualification Tests, including 75 verbal items in one subtest, 50 numerical items in the second, and 75 general-information items in the third. A multiple-regression analysis was carried out in seven separate samples and the weights which were derived were then used to predict the criterion in the original sample plus some of the other samples. Four colleges participated in the study. In three, both a male and a female sample was available, whereas in the fourth college only women were enrolled. Weights were crossvalidated only in samples of the same sex. Validity coefficients for each weighting and no weighting are presented in Table 3. Column headings refer to the school where the weights were derived; row headings refer to the school where the weights were applied.

Table 3

Crossvalidation of Multiple-Regression Weighting of Scores from Three Tests

Weights Applied			Unweighted	Weights Derived on College			
to College	Sex	*N*	Validity	A	B	C	D
A	M	449	.46	*.46*	.45	.43	
B	M	151	.51	.53	*.54*	.52	
C	M	217	.60	.59	.60	*.60*	
A	F	262	.59	*.59*	.55	.58	.58
B	F	169	.65	.66	*.68*	.65	.66
C	F	76	.52	.54	.49	*.56*	.52
D	F	107	.71	.71	.71	.69	*.71*

Note that in four of seven instances weighting did not improve validity at all, and that in the remaining three cases the increase was rather small. Of necessity the weights derived on a particular sample do at least a bit better in that sample than any other weights. But interestingly, the weights derived in one sample may do even better in another sample than they do in the original one. This of course reflects the fact that the validity is simply higher in some samples than in others and that it matters little what weights are used. Had each set of weights been derived on a *random* sample from the population of interest, *e.g.*, College A - Females, and crossvalidated on another random sample from the same population, we would expect the correlation to shrink rather than increase.

In a third study of interest (Booth, 1968) a multiple-regression technique was used to weight course grades in arriving at a final grade in two Naval Aviation Schools, the Aviation Officer Candidate School (AOCS) and Flight Preparation School (FP). All student naval aviators attend FP, as do naval flight officers. These groups are procured from AOCS or some other source. In this study the aim was to investigate whether a new set of weights would improve prediction of completion *vs.*non-completion of the training. Also considered was the possibility that subgroups might be used to derive special sets of weights and thereby improve prediction further. The subgroups of interest were formed by the intersection of two two-way classifications, *i.e.*, AOCS students *vs.* non-AOCS students, and student naval aviators *vs.* naval flight officers. Obviously, the first classification can be applied only to those in FP. The *n*'s were as follows: In AOCS, 839 students were student naval officers and 327 were naval flight officers. Of these, 812 and 303 went on to FP and formed the AOCS group. The non-AOCS group in FP contained 1122 student naval officers and 339 naval flight officers. Thus, the sample size in this study was sufficiently large that reasonably stable regression weights might be expected.

It was found that the new weights raised the correlation of final grade with the criterion from .207 in AOCS to .268, and from .304 in FP to .313, where the first figure is the validity under the old weighting scheme. When validity is computed separately for each subgroup, differences are revealed which are consistent regardless of which weighting scheme is used. The final grade is more valid for naval flight officers than for student naval officers in AOCS, and it is more valid for the AOCS students than for the non-AOCS students in FP. The use of special sets of weights for each of these four subgroups resulted in very slightly higher correlations, but not sufficiently higher to justify the difficulty involved in applying them.

It is interesting to contrast this study with the previous one where overall gains due to weighting were small. Of course in Booth's study weighting is not compared to natural random weighting but rather to a former method of weighting which is in fact highly correlated with the new one. In the previous study only three scores were combined in the composite, whereas in the present case there were eight course grades to be combined in AOCS and six in FP. With fewer variables in the composite, there should have been more opportunity for weighting to be effective in the first case than in the present study. What seems to lie behind the difference, however, is not the number of variables, but rather their average intercorrelation. In the first study the tests were moderately correlated with one another, whereas in the second study the course grades intercorrelated .19 on the average in AOCS and .34 in FP. This difference seems to be reflected also in the fact that the increase in the validity is smaller in the group where the intercorrelation is larger.

Perloff (1951) studied special procedures to reduce the shrinkage of validity coefficients when predictor weights based on one sample are applied to another. His results were somewhat equivocal.

The Weighting of Test Items

In this section we consider the case where the variables to be weighted are the items of a test. In the typical case the item itself is scored on a pass-fail dichotomy and then multiplied by the appropriate weight. From the earlier discussion of the factors which affect the effectiveness of weighting it would seem that tests consisting of a large number of items, perhaps on the order of 50 to 100, all or nearly all of which are positively correlated, are not likely to become much more valid or reliable under differential weighting, simply because the correlation between two such weighted tests will very likely approach unity. This is indeed the case. Yet numerous studies have been performed demonstrating empirically that this is true. In many cases, the conclusion to disregard weighting is based only on findings of high correlation between weighted and unweighted tests or between two differently weighted tests. As pointed out earlier, this still leaves open the possibility of differential validity. Yet even in those cases where validity coefficients are presented the gains attributable to weighting are so small as to be of no practical significance. The studies to be discussed in this section are typical of those which have been performed.

The arrival of the new-type or objective examination in the 1920's was accompanied by claims of objectivity in scoring which would result in fairer assignment of course grades and the like. There were opponents of the new tests, however, and some felt that in actuality the new tests were no more objective than the old ones. One such opponent was Corey, who in 1930 published a study which purported to demonstrate the element of subjectivity in new-type examinations. Corey asked six instructors to rate each of the 73 items of an objective test according to "its importance for a general knowledge of psychology." The ratings supplied by each instructor became the weights for the items. Corey scored all the examination papers without weights and with each of the six sets of weights. The judges' weighted test scores correlated from .836 to .960 with the unweighted totals, with the former figure being the more typical. Corey established arbitrary cut-offs and assigned letter grades to the six series of tests. He concluded that many of the students would receive very different grades depending on whose weights were used to score the test, thus demonstrating the subjectivity which lingered in the new-type test.

Corey's "experiment" is important because it is probably the only study which claims to show that weighting makes a difference. Unfortunately, as an experiment it is open to criticism on a number of grounds. In a follow-up study, Odell (1931) revealed that some of Corey's instructors had weighted certain items zero, thus eliminating them from the test! Odell approached the whole weighting problem is a more systematic way. First he compared several methods of weighting, including: (1) natural raw-score weighting; (2) weighting by the percentage of subjects answering correctly; (3) weighting by the percentage of subjects failing to answer correctly; (4) weighting by a random distribution of weights from 1 to 5; (5) weighting by a random distribution of weights from 1 to 10; (6) weighting by a second random distribution of weights from 1 to 5. With the exception of sets (2) and (3), which correlated -.62, all other sets of weights correlated near zero. When test scores were then calculated with these sets of weights the scores correlated in the .90's, much higher than in Corey's study. Odell then had instructors assign weights as in the Corey study. The three sets of weights thus obtained were moderately well correlated and again the test scores computed with the different sets of weights were almost perfectly correlated. Odell concluded that there was no evidence for the utility of weighting.

In both Corey's and Odell's study no information concerning the reliability or validity of the weighted and unweighted tests was presented. Guilford, Lovell, & Williams (1942) in a classic experiment compared weighted and unweighted scoring of a single multiple-choice test in terms of the effect on both reliability and validity. One hundred multiple-choice questions from an achievement examination were used in unweighted form as the criterion. Three "tests" were then composed of the first 20 items, the first 50 items, and the 100 items. Guilford's weight (see p.32) was then used to weight the items. The reliability of the tests was determined by the split-half reliability stepped up with the Spearman-Brown formula. The reliability and validity coefficients for the weighted and unweighted tests are presented in Table 4.

Table 4

Reliability and Validity of Weighted and Unweighted Tests

Number of Items	Reliability		Validity	
	Weighted	Unweighted	Weighted	Unweighted
20	.667	.649	.817	.793
50	.860	.844	.892	.901
100	.922	.899	.900	.924

Differences in reliability and validity for the weighted and unweighted tests are not significant. Guilford explains that the phi coefficients for these items and the range of the weights were both small. These facts might explain the failure of the weights to affect either reliability or validity. He also notes that since the validity coefficient in this case was a part-whole correlation for the unweighted tests, the spuriousness of this correlation may have obscured real differences. Yet attempting to derive an estimate of the correlation of the 100-item test without the spuriousness did not support this interpretation. Guilford's conclusion was that it was certainly not worth the trouble to weight the test items.

Other studies of item weighting have reached similar conclusions. Douglass & Spencer (1923) found weighted and unweighted tests to correlate .98, .99, .995, .996, .985, and .991. Holzinger reports a correlation of over .99 for weighted *vs.* unweighted items of a French achievement test (Holzinger, 1923). West (1924) found correlations ranging from .987 to .997 for weighted *vs.* unweighted comprehension tests. In addition, he reports correlations of .975, .956, .932, .966, .984, and .940 for six of the Army Alpha tests, weighted *vs.* unweighted. Peatman (1930), using Clark's Index of Validity to weight true-false items, found over a series of quizzes and a final exam that correlations ranged from .879 to .970 for the individual tests and that the correlation for all tests combined was .978. Ruch & Meyer (1931) found that weighting on the basis of difficulty did not raise validity and perhaps lowered reliability. Pothoff & Barnett (1932), in a study quite similar to that of Odell, found correlations of .965 to .987 between weighted and unweighted scores, when weights were based on teachers' opinions. Finally, Stalnaker (1938), in a study of weighting essay-type examinations, found correlations consistently on the order of .98 and .99 between weighted and unweighted versions of a number of examinations of the College Entrance Examination Board. Thus, it seems abundantly clear that weighting a given item of a test the same for all examinees simply does not affect the total score enough to be of practical significance. Although a great many of these studies report only the correlations between the weighted and unweighted scores, when the magnitude of such correlations is .98 or .99 it must be admitted that even if the small possible gain in validity allowed by such a correlation were to be expected, it would be too small to justify the extra amount of time and effort required to score the test using the weights. The utility of fixed item weighting seems to have long since been disproven.

There remain two possible hopes for effective differential weighting of item scores. One is Allan Birnbaum's work on a three-parameter logistic latent-trait model, reported in Lord & Novick (1968, ch. 17-20). Lord (1967) tried out this model with the Scholastic Aptitude Test Verbal scores of nearly 3000 examinees and reported his results cautiously but with guarded optimism.

The essence of Birnbaum's procedure is that it applies differential weights not only to items but also to various ability levels. His scoring produces the most improvement for the least able examinees, who throw noise into the system by guessing wildly at the more difficult items. In effect, Birnbaum's method seems to nullify such guessing by assigning small weights to items difficult for the examinees at that ability level, and larger weights to the easier items there.

Unfortunately, as Lord points out, Birnbaum's model applies only to data where there are no omitted responses, and Lord's comparison is with items of the Verbal Scholastic Aptitude Test scored either right or wrong, 1 or 0. The conclusions may not apply to Verbal SAT scored in the usual operational way with a "correction for chance." We know that giving -1/4 point for each Verbal SAT item marked incorrectly will tend to remove some of the effects of sheer guessing, thereby lessening the spurious intercorrelation of items and perhaps improving validity.

Lord also warns that his conclusions, which favor Birnbaum's procedure, depend on the adequacy of Birnbaum's mathematical model for describing Verbal SAT data. Despite these considerations and the complexity of the computations required, Birnbaum's approach seems promising enough to be investigated much further. It may use effectively a different kind of weighting, *i.e.*, by ability level, which is needed to go beyond the impasse clearly pointed out by Wilks (1938) and convincingly demonstrated by Stalnaker (1938) and others.

Cleary (1966) developed a model for multiple regression that allows individual differences to emerge empirically. This model effectively reduces the variance of errors of prediction, the weights obtained are stable over samples, and it appears that these weights have some stability over different sets of predictors. The model assigns to each person a different set of regression weights. It offers an empirical method of estimating whether prediction can be improved by deviating from the usual multiple-regression model and how many dimensions are required for maximum improvement. Her model goes beyond the situation considered by Wilks (1938), where there was just one set of weights, the same for each person, so it *might* provide a way to weight item scores in order to predict a criterion better than the nominally-equally-weighted item scores do. This moderated-linear-regression approach seems to be a possible alternative to Birnbaum's (1968) differential weighting of ability levels. Presumably, it could operate either with corrected-for-chance item scores or with uncorrected scores, whereas Birnbaum's procedure seems confined to the latter.

Also to be noted is Samejima's (1968) application of her graded-response model to multiple choice situations in an attempt to estimate latent ability. It did not prove successful for this particular case, but she promises further developments.

The Weighting of Item Responses

There are at least two very distinct situations where response weighting might be advantageous. First, when there is no correct answer to a question, as is the case in interest, attitude, and personality tests, the responses are usually weighted in order to differentiate between examinees. The methods of weighting such responses were considered in an earlier section. A second situation where response weighting might be profitable is in the academic achievement or aptitude test where partial information or misinformation is evaluated through differential weighting of the correct responses to each item. The second situation, although promising, has not been extensively investigated.

Probably no single test has been the subject of so many empirical investigations of weighting as the Strong Vocational Interest Blank. The method used by Strong to derive the weights for item responses was discussed earlier. It will be recalled that the size of the weight is related to the coefficient of the correlation between membership in the occupation or not, and choosing the response *vs.* not choosing it. This coefficient in turn is related to the size of the difference in the proportions of those choosing the response in the two populations. The larger the difference, the larger the coefficient. Responses which differentiate strongly between the two groups receive large weights relative to those which do not. In the original scoring system the weights ranged from 30 to -30. In the 1930 revision the range dropped to 4 to -4. More recently the adoption of unit weights 1, 0 and -1 has been advocated. This progressive collapse of the elaborate weighting system of the SVIB has resulted from a long series of experimental studies which demonstrated such slight reductions in predictive accuracy that it was concluded that the simpler weights were to be preferred to the more cumbersome ones.

The list of empirical studies of weighting the SVIB begins with the contention by Strong (1930) that the use of unit weights resulted in less differentiation between the occupations. In a series of experiments Dunlap and his associates claimed to have shown that the unit weights could in fact be substituted for the larger weights with only a small loss in accuracy (Dunlap, 1940; Peterson & Dunlap, 1941; Harper & Dunlap, 1942; Lester & Traxler,1942; Kogan & Gehlmann, 1942). The basic strategy in each of these studies is to score a sample of blanks with both unit and regular weights and then to use a multiple-regression equation to predict the fully weighted scores from the unit-weighted scores. The regression weights are then used to predict the weighted scores in a cross-validation group, and the correlation between predicted and actual scores in computed. It is usually in the mid to high .90s . Since the SVIB is used as the basis of vocational counseling, an important question is to what extent is the letter-grade designation upon which counseling is based affected by the change in scoring procedure? Thus, in each study, the letter grades are assigned on the basis of predicted and actual scores and the percentages of correct classifications is reported, with special attention to the shift of the B+ scores to B, a change which corresponds to a failure to recommend the ocucupation. Usually, the critical shift occurs in about 3.5% of all cases. Strong (1945),in an extensive review of this research, claimed that not only the highest scores on the blank were to be stressed, but the entire pattern, and that additional changes in the scores might noticeably affect this pattern. Strong maintained even in 1964, when the unit weights were finally adopted, that unit scoring reduced validity. However, under the considerable pressure put forth by others, the SVIB finally acquired unit weights (Strong, Campbell, Berdie & Clark,1964).

Essentially, similar findings have been reported in research with the Bernreuter Personality Inventory (Bennett, 1938; Kempfer, 1944; McClelland, 1944, 1947). Here also a small loss of accuracy is suffered when diminished weights are used.

Until fairly recently, the possibility of differentially weighting the incorrect responses of an achievement examination had not been considered in the literature. It has long been assumed that on a multiple-choice examination the conventional correction-for-guessing formula provides a reasonable means of deducting from the number-correct score the proportion of those correct items which are the result of random guessing. Formula scoring is, in at least one sense, response weighting. If the conventional formula is used, where Score = Right - $[1/(k - 1)]$ Wrong, this is equivalent to assigning a weight of +1.00 to each correct response, $-1/(k - 1)$ to every incorrect response, and 0 to an omitted item. The subject's score is then the algebraic sum of the responses he selects. Some investigatore have preferred to empirically derive the best weight for the incorrect response via some technique such as multiple regression (*e.g.*,see Thurstone, 1919; Brinkley, 1924; Staffelbach, 1930; Dailey,1947). However, formula scoring, regardless of how the formula is derived, is not *differential* weighting of distractors other than "omit." The subject's score depends on the number of correct responses, the number of incorrect responses, and the number of omissions, but it is not affected by *which* incorrect response is chosen on a particular item.

If each examinee omits the same number ot items, formula scoring us not needed. (See Stanley, 1954) For theoretical increases in validity (usually slight) that might be gained by formula scoring, see Lord(1963).

The first step in the direction of differential weighting of incorrect responses was made by Nedelsky (1954). He hypothesized that information might be added to the conventional Rights score by penalizing students for choosing a response which was so grossly incorrect as to be attractive only to an F-student. Nedelsky had experts read his test questions and indicate which options for a given item, if any, fit this description. The experts' judgements were the basis for the designation of certain incorrect responses to the questions as *F*- responses. Some items had no such responses, others had more than one. The test was given to 651 students and each received three scores: (1) a rights score; (2) an *F*-score (3) a composite score computed on the formula $R - F/f$, where f is the average number of F responses available per item. Of the 651 who took the exam, all receiving a D or an F by standard scoring, plus a representative sample of those receiving an A,B, or C, 306 in all, were rescored for *F*-score and for the composite. The reliability of the rights score was estimated as .81, of the *F*-score,.63, and of the composite,.84. However, when these figures are computed separately for the ABC group and the DF group they become respectively: ABC: rights, .69, *F*-score,.46, composite, .71; DF: rights,-.16, *F*-score,.42, composite,.26. Thus for the poorer students the *F*-score is the most reliable score. Accentuating this finding is the fact that when the reliability is oomputed only on those items having *F*-responses, and only for the lowest 15% of the entire 651 subjects, the figure for the *F*-score rises to .45. What is particularly interesting in this study is that for both groups of subjects the composite score is more reliable than the rights-only score.

If Nedelsky's study may be looked on as a significant first step in the direction of differential response weighting, the later study of Davis & Fifer (1959) may be considered a significant second one. These authors note that conventional rights-only scores, as well as formula scores, do not permit differentiation among examinees with respect to the type of distracters they select. The student who consistently chooses incorrect responses which are most nearly correct receives the same penalty as the student who chooses the same number of incorrect responses, but whose choices reflect very little information at all. If it is possible to differentiate among incorrect alternatives with respect to their degree of incorrectness, then it might be worthwhile to weight these alternatives differentially.

From a pool of 300 arithmetic-reasoning problems two tests of 45 items each were contructed and designated test 5022 and test 5023. An additional five problems testing computational facility were also included as a sort of handicap to eliminate unwanted variance from this source in the total score. *A priori* weights were derived via ratings given to each response. Two mathematicians were instructed to rate each response option on a seven-point scale from -3 to +3 according to the relative amount of arithmetic reasoning, *i.e.*, correct reasoning , displayed by an examinee marking that option. In general such weights were positive for the correct response and negative for the incorrect ones.

Empirical weights were derived via the correlation between marking the option *vs*. not marking it and the criterion score on both tests 5022 and 5023. These weights were then transformed to the range -3 to +3. The authors do not give in detail their reasons for using correlation coefficients except to say that these are approximations to the appropriate multiple-regression weights. (Davis, 1959, is more explicit.) Since the responses are categories rather than variables, the weights seem somewhat less appropriate than Guttman's (1941) criterion weights, though considerably easier to secure. However, since the reliability of the weights based on the correlation coefficients was only moderately high, the empirical weights were modified in terms of the *a priori* weights to arrive at a final set of weights for each item. Two studies were then carried out.

A sample of examinees from the larger group which had taken the tests, and which did not include any of those in the sample on which the weights were derived, was used to estimate the reliability of the two scoring procedures. Two raw scores and two weighted scores were available for each subject. Since 5022 and 5023 were considered parallel tests, their correlation was used to estimate the reliability of either one. Unweighted scores were found to correlate .6836, weighted scores .7632. The difference between these two correlations is highly significant by Fisher's *z*-transformation. It was estimated that the increase in reliability was equivalent to that which would be expected by the Spearman-Brown formula if the test were lengthened from 45 items to 67 items and scored conventionally. It is pointed out that the increase in reliability is not to be attributed to the fact that the *correct* choices were differentially weighted since it has long been known that such differential weighting is not effective for long tests.

In a second study an attempt was made to determine whether the new scoring procedure would increase the validity of the test. Two criteria were used, teacher's ratings and the score on a free-response form of the same tests. Very briefly, it was found that for 251 subjects who took one test in free form and the other in multiple-choice form, the validity using these criterion variables was not different. The authors conclude that the variance introduced into the total score increased the proportion of "true" variance, but that the new variance had the same concurrent validity as the original.

A slight exception may be taken to this conclusion. If the new variance were as valid as the original variance, then the increased reliability of the test should have resulted in a concomitant increase in the validity, much as increasing the length of a test through the addition of items of comparable reliability and validity does. However, before meaningful generalizations can be made concerning the nature of the variance which is added to the tc tal test score through this type of differential weighting it will be necessary to explore more carefully the composition of the total-test-score variance. Aiken (1967) has presented formulas for the maximum total variance of the test score, but some of his assumptions do not seem fully justified. More work on this aspect of the problem is needed.

A rather novel recent study is that of Jacobs and Vandeventer (1968), where "the notion of facet analysis provided a systematic method for *a priori* ordering of the distractors on the Coloured Progressive Matrices test as to degree of correctness. A score based on type of distractor chosen was shown to have a moderate degree of test-retest reliability, concurrent and predictive validity, and cross-cultural applicability."

Variable Weighting Methods

In the beginning of this paper it was stated that throughout most of the discussions the weighting methods considered would be fixed methods. Fixed, of course, refers to the fact that the weights are determined in advance of scoring the subject's paper and that a definite weight is attached to each item or response. The subject's score on the item or response is determined by a binary outcome, *viz.*, correct *vs.* incorrect for the weighting of items, and chosen *vs.* not chosen for the weighting of responses.

Modification of the Mode of Response.

Recently, however, a somewhat different approach has been investigated. Since the very beginning of the objective test movement there has been considerable concern over the effects of guessing on the reliability and validity of multiple-choice and true-false tests. Most often when testers were sufficiently concerned over the effects of guessing to attempt to correct for it, they relied on some form of correction-for-guessing formula which subtracted a percentage of the incorrect responses from the correct ones. The traditional formula, $R - [1/(k - 1)]W$, is based on the assumption that if a subject does not know the answer to a question he guesses randomly among the k options.

Admittedly, this assumption is never satisfied in practice. Response alternatives differ in attractiveness, as the differential popularity of incorrect options indicates. Moreover, the assumption asserts that information comes in two states, certainty and complete ignorance. The existence of both partial information and misinformation is thus denied. However, the usefulness of the traditional formula has served to perpetuate it.

In the last section the door was opened for partial information to reveal itself through the differential weighting of the distracters. It was implicit in the scoring scheme that the incorrect options could serve to identify the existence of partial information. In actuality, the empirical method of deriving option weights does not ensure that the more heavily weighted options are in fact more nearly correct, but merely that on the average the total scores of subjects choosing the more heavily weighted option are higher than those of subjects choosing another option on the same item. In this type of weighting scheme it is the options themselves which bear the burden of differentiating between the subjects with respect to partial information.

There is an alternative approach, however. If we assume that the subject has some information concerning the correctness of the several options, instead of assuming that correct choices are made out of certainty and incorrect ones out of guessing, we may, through appropriate response techniques and scoring procedures, lead the subject to reveal much more precisely the actual state of his information concerning all of the options.

In each of the methods of variable weighting considered below, the mode of response to the individual test item has been altered from that of the traditional multiple-choice item. Scoring is not carried out on the basis of a weighting of individual items or item options.

Elimination of Response Alternatives

In two fairly recent studies the mode of response was altered by having subjects cross out options. Dressel & Schmid (1953) compared the conventional multiple-choice paradigm with one in which the subjects were instructed to cross out alternatives until they were certain that they had included the correct answer among the alternatives marked. Each incorrect mark was scored as - 1/4 point. Thus, marking all alternatives except the correct one resulted in the maximum negative score and marking only the correct choice resulted in the maximum positive score. This scoring method and response technique was found to give a reliability of .67 as compared with a reliability of .70 for the conventional procedure.

Coombs, Milholland & Womer (1956) performed the complementary experiment where subjects were instructed to cross out the incorrect alternatives, taking care not to mark the correct alternative. Each incorrect option eliminated received a score of +1 and if the correct alternative was marked it was scored - $(k - 1)$, where k is the number of choices for the question. Thus, if r alternatives were marked, the score was $+r$ if the correct alternative was not marked, and $(r - k)$ if it was. Marking all alternatives or no alternatives resulted in a score of zero. There was evidence to indicate that this method of scoring resulted in a gain of reliability equivalent to that to be expected by increasing the length of the test 20%.

These techniques are reminiscent of the Troyer-Angell punchboard invented two decades ago and sold by Science Research Associates. The punchboard was a device on which the subject punched out his choice and if it was correct a red dot would appear in the hole. If the dot did not appear the subject had to choose another response until the dot did appear. Thus, the subject received immediate feedback on the correctness of his choices, learning while being tested. He was then scored on the basis of the number of punches needed to reveal the dot: 0, -3, -4, -6, -7, for correct answer on the first through 5th responses respectively. When two groups of subjects used the punchboard for an entire semester and did not use it respectively, the difference between the groups, favoring the users, increased during the semester from zero to a value approaching statistical significance. (See Jones & Sawyer, 1949).

Confidcnce Weighting

A second method of assessing partial information which has shown some promise is that of having students assign confidence weights to the various alternatives. This procedure has its historical antecedents in the confidence weighting of true-false tests (Hevner, 1932; Soderquist, 1936). More recently it has been tried with multiple-choice tests. Dressel & Schmid (1953) also included this as a scheme in their study. They had subjects choose one alternative and then assign a confidence weighting from 1 to 4 in accordance with their degree of certainty regarding the correctness of their choice. The weight was scored as positive if the item choice was correct and negative if incorrect. They report a reliability of .73 for this case as opposed to the .70 for the conventional case. Also see Merwin (1959) for a theoretical analysis of the effects of ranking options according to preference for them.

Subjective Probabilities

Recently, from two different sources (Shuford, Albert, & Massengill, 1966; de Finetti, 1965) have come suggestions concerning the assignment of probabilities to each of the response alternatives for a single question. The most interesting characteristic of these procedures is that, under what has been termed admissible probability measurement, the scoring system is so devised that the examinee can maximize his expected score on the test if and only if he reports as accurately as he can the distribution of his subjective probabilities over the various response options. Not all scoring methods achieve this, and considerable mathematics has been devoted to illustrating permissible schemes.

In the Shuford, Albert, & Massengill procedure it is assumed initially that the examinee's state of knowledge concerning a multiple-choice item may be expressed as a distribution of probabilities over the response options. Since probability distributions of this kind are not "wired into" the cognitive system, it is assumed that examinees are able to convert their degrees of confidence in the various options into a probability distribution having the property that the sum of the probabilities over all response options is 1.00.

It should be pointed out that since such probabilities are subjective and, presumably determined by the *relative* degrees of confidence which the examinee places in the correctness of the various options, there is no guarantee that identical probability distributions for two subjects represent the same absolute degrees of confidence in the options taken individually. The probabilities are ipsative measures. If, for example, the examinee assigned equal probability to each option, he might do so out of complete ignorance or out of conflicting misinformation which gave him rather high confidence in each option considered singly. The assigned probabilities must be seen as measures of *relative* confidence.

Once the examinee's probability distribution is known for a multiple-choice item, there are numerous scoring techniques which may be used to determine the item score. Shuford, Albert, & Massengill have discussed the necessary and sufficient mathematical properties of scoring schemes which have the property of allowing the subject to maximize his expected score if and only if he reports his "true" subjective probabilities. For items with more than two options it is not possible to find such a scheme which is dependent only on the probability assigned to the correct answer. Most schemes thus involve the distribution of confidence over the incorrect options and most are symmetric in that the item score does not change when the probabilities assigned to the incorrect options are permuted. Thus there is usually no differential penalty for assigning high confidence to one incorrect option over another.

Shuford and Massengill, in a series of technical reports (Shuford, 1967; Shuford & Massengill, 1967; Massengill, 1967), have expressed great enthusiasm and optimism concerning the potential of admissible probability measurement for eliminating the effects of guessing on multiple-choice tests. They have demonstrated mathematically that the elimination of guessing, which is purportedly accomplished by admissible probability measurement, can *theoretically* provide quite substantial gains in both reliability and validity. These maximum gains, however, are determined with reference to the reliability and validity of a multiple-choice test in which: (1) the level of guessing, *i.e.*, the probability of being correct given that guessing occurs, is at a maximum level of .50; (2) all examinees guess when they do not "know" the correct answer; and (3) the test is scored as the number of items correct. For example, they demonstrate that for a test so difficult that no examinee knows the answer to any question, more guessing by some examinees than by others results in a spurious reliability over a large number of items or examinees. It is easy to show, however, that insofar as the conventional correction-for-guessing formula accurately reflects the *average* level of guessing, the expected value of the reliability of formula scores is zero here. Although the elimination of guessing may plausibly increase reliability and validity, the size of these gains will be considerably smaller than the maximum possible gains, since in actual practice the effects of guessing are not nearly so devastating as they might be, particularly when a correction-for-guessing formula is used. (Also see Lord, 1963.)

Although the Shuford, Albert , & Massengill admissible-probability-measurement procedure requires the examinee to report his subjective probabilities directly, other response methods are possible. De Finetti (1965) has discussed a number of response schemes which provide various degrees of information concerning the subject's probability distribution. It is assumed that the probability distribution is directly available to the examinee and that he may use the distribution to determine his response so as to maximize his expected score on the item. It is paradoxical that although the actual responses made by the student seem superficially to be simpler than assigning probabilities directly, the optimal strategy required of the examinee may take the form of a very complicated rule. For example, assume that the subject is to respond by crossing out the incorrect options on a multiple-choice item, with the number of options eliminated left to the student to decide. (This is the response method of Coombs *et al.*) If r is the number of options, and k is the number of options crossed out, the score is determined by the formula $1/(r - k)$, and made negative if the correct answer is crossed out and positive if it is not. How many options should the subject cross out given his probability distribution? Let the subject rank the options such that p_1 is the largest probability and p_h is the probability assigned to the hth option, $h = 1,\ldots,r$. The rule for maximizing the expected score on this item may then be stated: "Cross out alternatives until the probability p_h' of the $(r - h)$ alternatives already crossed out plus that of the next one, p_h^*, when multiplied by the number h of those still left, does not attain .5" (de Finetti, 1965, p.98).

Thus, although the response of crossing out alternatives seems relatively undemanding, the strategy which will allow the subject to be consistent and maximize his expected score is more than a little complex!

The success of this type of testing procedure would be critically dependent on the ability of subjects to effectively utilize optimum strategies. The problem of whether or not all subjects are equally capable of learning to use such strategies is a very real one. Also raised is the problem of the differential risk-taking propensities of different subjects. Despite the fact that risk taking must *in the long run* reduce the expected score, the score on a single test can be altered significantly by a lucky guess. Winkler (1967a, 1967b, 1967c) has discussed these aspects of subjective probability measurement.

It will be recalled that most admissible probability measurement or subjective probability measurement procedures are symmetrical and do not take into account characteristics of specific distracters. In these scoring systems involving probabilities *per se*, high concentration of confidence in a single distracter results in a lower score than an equal distribution of that same amount of confidence over all distracters. It might, however, be possible to differentially weight distracters and incorporate such weights into the scoring scheme. Although some type of criterion keying of options could probably be incorporated while maintaining the admissible probability of the scoring, it is likely that the optimal strategy would then become considerably more complicated. Empirical work on the use of admissible probability measurement and differential option weighting, both separately and in conjunction, is undoubtedly forthcoming since so much theoretical interest in these proposals has been aroused.

References

Aiken, L.R. Effect on test score variance of differential weighting of item responses. *Psychological Reports*, 1967, *21*, 585-590.

Anastasi, A., Meade, M.J., & Schneiders, A.A. *The validation of a biographical inventory as a predictor of college success.* N.Y.: College Entrance Examination Board, 1960.

Anderson, R.G. A critical examination of test scoring methods. *Archives of Psychology*, No.80, 1925.

Bennett, G.K. A simplified scoring method for the Bernreuter Personality Inventory. *Journal of Applied Psychology*, 1938, *22*, 390-394.

Bingham, W.V. Reliability, validity, and dependability. *Repr. & Cir. Ser. Person. Res. Fed.*, No.24, 1932.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. Part 5 (Chs. 17-20, pp. 395-479) in F.M. Lord & M.R. Novick, *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Booth, R.F. Optimal weighting of course grades for two naval air training schools. *NAMI-1039*. Pensacola, Fla.: Naval Aerospace Medical Institute, 3 May 1968.

Bovee, A.G., Holzinger, K.J., & Morrison, H.C. The construction of tests for the measurement of certain achievements in French. *Supplementary Educational Monographs*, No.26, University of Chicago, 1925, 109-136.

Brinkley, S.G. Values of new-type examinations in the high school. *Contributions to Education*, No.161. N.Y.: Teacher's College, Columbia University, 1924.

Burt, C. The influence of differential weighting. *British Journal of Psychology, Statistical Section*, 1950, *3*, 105-125.

Clark, E.L. A method of evaluating the units of a test. *Journal of Educational Psychology*, 1928, *19*, 263-265.

Cleary, T.A. An individual differences model for multiple regression. *Psychometrika*, 1966, *31*, 215-224.

Cliff, R. The effect of unlike distributions on the weights of variables. *Educational and Psychological Measurement*, 1960, *20*, 305-310.

Congdon, N.A. New weights for the responses in the Heilman Personal Data Scale. *Journal of Educational Psychology*, 1941, *32*, 214-219.

Coombs, C.H., Milholland, J.E., & Womer, F.B. The assessment of partial knowledge. *Educational and Psychological Measurement*, 1956, *16*, 13-37.

Corey, S.M. The effect of weighting exercises in a new-type examination. *Journal of Educational Psychology*, 1930, *21*, 383-385.

Cowdery, K.M. An evaluation of the expressed attitudes of members of three professions. Unpublished doctoral dissertation, Stanford University, 1925.

Cureton, E.E. Validity, reliability, and baloney. *Educational and Psychological Measurement*, 1950, *10*, 94-96.

Cureton, E.E. Approximate linear restraints and best predictor weights. *Educational and Psychological Measurement*, 1951, *11*, 12-[illegible].

Dailey, J.T. Techniques for estimating the optional weight of the "wrong" in scoring printed tests. *American Psychologist*, 1947, *2*, 310-311. (Abs.)

Davis, F.B. Estimation and use of scoring weights for each choice in multiple choice test items. *Educational and Psychological Measurement*, 1959, *19*, 291-298.

Davis, F.B., & Fifer, G. The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, 1959, *19*, 159-170.

Douglass, H.R., & Spencer, P.L. Is it necessary to weight exercises in standard tests? *Journal of Educational Psychology*, 1923, *14*, 109-112.

Dressel, P.L., & Schmid, P. Some modifications of the multiple choice item. *Educational and Psychological Measurement*, 1953, *13*, 574-595.

Dunlap, J.W. Simplification of the scoring of the Strong Vocational Interest Blank. *Psychological Bulletin*, 1940, *37*, 450. (Abs.)

Dunnette, M.D., & Hogatt, A.C. Deriving a composite score from several measures of the same attribute. *Educational and Psychological Measurement*, 1957, *17*, 423-434.

Edgerton, H.A., & Kolbe, L.E. The method of minimum variation for the combination of criteria. *Psychometrika*, 1936, *1*, 183-187.

de Finetti, B. Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 1965, *18*, 87-123.

Gage, N.L. Logical *vs*. empirical scoring keys: the case of the MTAI. *Journal of Educational Psychology*, 1957, *48*, 213-216.

Giles, G.R. A new interests test. *Journal of Educational Psychology*, 1936, *27*, 527-536.

Glass, G.V, & Maguire, T.O. Abuses of factor scores. *American Educational Research Journal*, 1966, *3*, 297-304.

Guilford, J.P. A simple scoring weight for test items and its reliability. *Psychometrika*, 1941, *6*, 367-374.

Guilford, J.P. *Psychometric methods.* N.Y.: McGraw-Hill, 1954.

Guilford, J.P., Lovell, C., & Williams, R.M. Completely weighted *versus* unweighted scoring in an achievement examination. *Educational and Psychological Measurement*, 1942, *2*, 15-21.

Gulliksen, H. *Theory of mental tests.* N.Y.: Wiley, 1950.

Guttman, L. An outline of the statistical theory of prediction. In Paul Horst, *The prediction of personal adjustment.* N.Y.: Social Science Research Council, 1941.

Harper, B.D., & Dunlap, J.W. Derivation and application of a unit scoring system for the Strong Vocational Interest Blank for Women. *Psychometrika*, 1942, *7*, 289-295.

Harris, C.W. On factors and factor scores. *Psychometrika*, 1967, *32*, 363-379.

Hevner, K.A. A method of correcting for guessing in true-false tests and empirical evidence in support of it. *Journal of Social Psychology*, 1932, *3*, 359-362.

Holzinger, K.J. An analysis of the errors in mental measurement. *Journal of Educational Psychology*, 1923, *14*, 278-288.

Horst, P. Obtaining a composite measure from a number of different measures of the same attribute. *Psychometrika*, 1936, *1*, 53-60.

Hotelling, H. The most predictable criterion. *Journal of Educational Psychology*, 1935, *26*, 139-142.

Jacobs, P.I., & Vandeventer, M. Information in wrong responses. *Research Bulletin* 68-25, *Educational Testing Service*, June 1968.

Jones, H.L., & Sawyer, M.O. New evaluation instrument. *Journal of Educational Research*, 1949, *42*, 381-385.

Jurgensen, C.E. Item weights in employee rating scales. *Journal of Applied Psychology*, 1955, *39*, 305-307.

Kaiser, H.F. Uncorrelated linear composites maximally related to a complex of correlated observations. *Educational and Psychological Measurement*, 1967, *27*, 3-6.

Kelley, T.L. *Statistical method.* N.Y.: Macmillan, 1923.

Kelley, T.L. The scoring of alternative responses with reference to some criterion. *Journal of Educational Psychology*, 1934, *25*, 504-510.

Kempfer, H. Simplifying the scoring technique of the Bernreuter Personality Inventory. *Journal of Applied Psychology*, 1944, *28*, 412-413.

Kogan, L., & Gehlmann, F. Validation of the simplified method for scoring the Strong Vocational Interest Blank for Men. *Journal of Educational Psychology*, 1942, *33*, 317-320.

Kuder, G.F. *Kuder Preference Record.* Chicago: Science Research Associates, 1934.

Lawshe, C.H., & Harris, D.H. The method of reciprocal averages in weighting personnel data. *Educational and Psychological Measurement*, 1958, *18*, 331-336.

Lawshe, C.H., & Schucker, R.E. The relative efficiency of four test weighting methods in multiple prediction. *Educational and Psychological Measurement*, 1959, *19*, 103-114.

Lester, H., & Traxler, A.E. Simplified method for scoring the Strong Vocational Interest Blank applied to a secondary-school group. *Journal of Educational Psychology*, 1942, *33*, 628-631.

Lord, F.M. Formula scoring and validity. *Educational and Psychological Measurement*, 1963, *23*, 663-672.

Lord, F.M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. *Research Bulletin* 67-34, *Educational Testing Service*, August 1967.

Lord, F.M., & Novick, M.R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley, 1968.

Massengill, H.E., & Shuford, E.H., Jr. What pupils and teachers should know about guessing. *Technical Report* SMC R-7, *Shuford-Massengill Corporation*, Lexington, Mass., May 1967.

McClelland, D.C. Simplified scoring of the Bernreuter Personality Inventory. *Journal of Applied Psychology*, 1944, *28*, 414-419.

McClelland, D.C. Further application of simplified scoring of the Bernreuter Personality Inventory. *Journal of Applied Psychology*, 1947, *31*, 182-188.

McCornack, R.L. A criticism of studies comparing item weighting methods. *Journal of Applied Psychology*, 1956, *40*, 343-345.

Merwin, J.C. Rational and mathematical relationships of six scoring procedures applicable to three-choice items. *Journal of Educational Psychology*, 1959, *50*, 153-161.

Mosier, C.I. On the reliability of a weighted composite. *Psychometrika*, 1943, *8*, 161-168.

Nedelsky, L. Ability to avoid gross error as a measure of achievement. *Educational and Psychological Measurement*, 1954, *14*, 459-472.

Odell, C.W. Further data concerning the effect of weighting exercises in new-type examinations. *Journal of Educational Psychology*, 1931, *22*, 700-704.

Peatman, J.G. The influence of weighted true-false test scores on grades. *Journal of Educational Psychology*, 1930, *21*, 143-147.

Peel, E.A. A short method for calculating maximum battery reliability. *Nature, London*, 1947, *159*, 816-817.

Peel, E.A. Prediction of a complex criterion and battery reliability. *British Journal of Psychology, Statistical Section*, 1948, *1*, 84-94.

Perloff, R. Using trend-fitting predictor weights to improve cross-validation. Unpublished doctoral dissertation, The Ohio State University, 1951.

Peterson, B.M., & Dunlap, J.W. A simplified method for scoring the Strong Vocational Interest Blank. *Journal of Consulting Psychology*, 1941, 5, 269-274.

Pintner, R. A standardization and weighting of two hundred analogies. *Journal of Applied Psychology*, 1920, *4*, 263-273.

Pothoff, E.F., & Barnett, N.E. A comparison of marks based upon weighted and unweighted items in a new-type examination. *Journal of Educational Psychology*, 1932, *23*, 92-98.

Ream, M.J. *Ability to sell.* Baltimore: Williams & Wilkins, 1924.

Richardson, M.W. The combination of measures. In Paul Horst, *The prediction of personal adjustment.* N.Y.: Social Science Research Council, 1941.

Ruch, G.M., & Meyer, S.H. Comparative merits of physics tests. *School Science and Mathematics*, 1931, *31*, 676-680.

Ryans, D.G. An analysis and comparison of certain techniques for weighting criterion data. *Educational and Psychological Measurement*, 1954, *14*, 449-458.

Samejima, F. Application of the graded response model to the nominal response and multiple-choice situations. *Research Report* No.63, *L.L. Thurstone Psychometric Laboratory, University of North Carolina,* July 1968.

Scates, D.E., & Fauntleroy, V. The effect of weights on certain index numbers. *Journal of Experimental Education*, 1938, *6*, 282-306.

Scates, D.E., & Noffsinger, F.R. Factors which influence the effectiveness of weighting. *Journal of Educational Research*, 1931, *24*, 280-285.

Shiba, S. A method for scoring multicategory items. *Japanese Psychological Research*, 1965, *7*, 75-79.

Shuford, E.H., Jr. How to shorten a test and increase its reliability and validity. *Technical Report* SMC R-9, *Shuford-Massengill Corporation*, Lexington, Mass., May 1967.

Shuford, E.H., Jr., Albert, A., & Massengill, H.E. Admissible probability measurement procedures. *Psychometrika,* 1966, *31*, 125-145.

Shuford, E.H., Jr., & Massengill, H.E. The relative efficiences of five instructional strategies. *Technical Report* SMC R-8, *Shuford-Massengill Corporation*, Lexington, Mass., June 1967.

Simpson, R.G. Weighting items for multiple response differentiation: an exploratory study for new methods. Unpublished M.S. thesis, Western Reserve University, 1951.

Soderquist, H.O. A new method of weighting scores in a true-false test. *Journal of Educational Research*, 1936, *30*, 290-292.

Staffelbach, E.H. Weighting responses in true-false examinations. *Journal of Educational Psychology*, 1930, *21*, 136-139.

Stalnaker, J.M. Weighting questions in the essay-type examination. *Journal of Educational Psychology*, 1938, *29*, 481-490.

Stanley, J.C. "Psychological" correction for chance. *Journal of Experimental Education*, 1954, *22*, 297-298.

Stanley, J.C., & Wang, M.D. Weighting test items and test-item options, an overview of the analytical and empirical literature. Mimeographed, August 1968.

Stanley, J.C., & Wang, M.D. Restrictions on the possible values of r_{12} given r_{13} and r_{23}. *Educational and Psychological Measurement*, 1969 (in press).

Stromgren, B. On certain mathematical problems connected with the determination of anthropometrical and diagnostical indices. *Acta psychiat. Kbh.*, 1946, *21*, 747-752.

Strong, E.K., Jr. Procedure for scoring an interest test. *Psychological Clinic*, 1930, *19*, 63-72.

Strong, E.K., Jr. *Vocational interests of men and women.* Stanford: Stanford University Press, 1943.

Strong, E.K., Jr. Weighted versus unit scales. *Journal of Educational Psychology*, 1945, *36*, 193-216.

Strong, E.K., Jr., Campbell, D.P., Berdie, R.F., & Clark, K.E. Proposed scoring changes for the Strong Vocational Interest Blank. *Journal of Applied Psychology*, 1964, *48*, 75-80.

Thomson, G.H. Weighting for battery reliability and prediction. *British Journal of Psychology*, 1940, *30*, 357-366.

Thurstone, L.L. A scoring method for mental tests. *Psychological Bulletin*, 1919, *16*, 235-240.

Tiffin, J., & Musser, W. Weighting merit rating systems. *Journal of Applied Psychology*, 1942, *26*, 575-583.

Wesman, A.G., & Bennett, G.K. Multiple regression *vs.* simple addition of scores in prediction of college grades. *Educational and Psychological Measurement*, 1959, *19*, 243-246.

West, P.V. The significance of weighted scores. *Journal of Educational Psychology*, 1924, *15*, 302-308.

Wherry, R.J. Maximal weighting of qualitative data. *Psychometrika*, 1944, *9*, 263-266.

Wilks, S.S. Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, 1938, *3*, 23-40.

Winkler, R.L. The assessment of prior distributions in Bayesian analysis. *Journal of the American Statistical Association*, 1967, *62*, 776-800. (a)

Winkler, R.L. The quantification of judgment: some methodological suggestions. *Journal of the American Statistical Association*, 1967, *62*, 1105-1120. (b)

Winkler, R.L. The quantification of judgment: some experimental results. *Proceedings of the American Statistical Association*, 1967, 386-395. (c)

Wolins, L. The use of multiple regression procedures when the predictor variables are psychological tests. *Educational and Psychological Measurement*, 1967, *27*, 821-827.

Wright, W.W. The development and use of a composite achievement test. *Indiana University School of Education Bulletin*, No.5, 1929.

Yerkes, R.M., Bridges, J.W., & Hardwick, R. *A point scale for measuring mental ability*. Baltimore: Warwick & York, Inc., 1915.